Structural Studies and Molecular Modelling of Alpha-2u-globulin

by

Paul David Adams

Ph.D. Thesis

University of Edinburgh 1992

Declaration

I declare that this thesis was composed by me, that the work of which it is a record was done by me, except where stated in the thesis. This work has not been accepted elsewhere in any previous application for a degree. All of the sources of information have been acknowledged.

Acknowledgements

I would like to thank both my supervisors, Dr. Lindsay Sawyer and Dr. Ted Lock, for their interest in my work during this project. I would also like to thank the following people:

- Peter Phillips for expert help in purifying MUP at ICI.
- Stella Fawcett for growing crystals.
- Paul Taylor for computing help.
- Sandy Blake for help in the small molecule work.
- The Edinburgh Parallel Computing Centre for use of facilities.
- Tim and Linda for being there.
- and finally to Elspeth, Alan, Mary, Joao, and all others from room 302 who made the work so much fun.

The financial support of both the Science and Engineering Research Council, and ICI is gratefully acknowledged.

Abstract

It has been observed that certain small hydrocarbon molecules produce an increased rate of kidney damage (nephropathy) and incidence of renal carcinomas in male rats. Both *d*-limonene and 2,4,4-trimethylpentane are seen to be active nephropathic agents. Human exposure to these chemicals is widespread as *d*-limonene is used as a lemon flavouring in some lemonade drinks, and 2,4,4-trimethylpentane is a hydrocarbon from petroleum. The possible nephropathic effects of such chemicals in humans is cause of great concern. On the basis of biochemical research a model has been proposed for the mechanism by which these chemicals cause kidney damage in male rats. A non-covalent association between the chemical and the urinary protein alpha-2u-globulin (a2u) is suggested. This association reduces the susceptibility of the protein to lysosomal proteolysis. The susequent accumulation of undigested protein in the lysosome eventually leads to cell death. Knowledge of the atomic structure of the protein is needed to understand how these chemicals bind and how this binding could affect proteolytic sensitivity. The work presented attempts to determine this atomic structure by direct crystallographic methods and by indirect homology modelling with the known structures of related proteins.

Crystals of a2u were grown, from protein purified to one molecular weight species, using the hanging drop method with ammonium sulphate as the precipitant. These crystals were small making X-ray analysis possible only with a synchrotron radiation source. The data collected indicated crystal twinning, which made interpretation of the data impossible. Efforts to improve crystal quality with narrower precipitant ranges and stabilising additives were unsuccessful. A homologous protein from mouse urine, major urinary protein (MUP), was purified using gel filtration followed by chromatofocusing and ion exchange chromatography. Crystallisation trials with this purified protein produced small crystals with the hanging drop method using either ammonium sulphate or ethanol as the precipitant. Crystals were both too small and temperature sensitive for X-ray analysis.

Both a2u and MUP are members of a family of related proteins, the lipocalycins. The structures of three of the members of the family were available (human retinol binding protein, bovine beta lactoglobulin, and tobbaco hornworm insecticyanin). It was possible to model the structure of a2u using these three crystallographically determined structures. Four different methods were used, the results of which were compared. The best model was selected on the basis of several criteria and was compared to the recently available structure of MUP. In all methods extensive energy minimisation in the presence of many solvent molecules was performed.

The extensive use of computational resource for energy minimisation and molecular dynamics calculations was prohibitive. Therefore, a parallel implementation of the GROMOS87 forcefield was written to run on a Meiko Computing Surface. The program was tested on a maximum of 132 T800 transputer nodes. This program was seen to give a 30 fold increase in performance, compared to the serial version running on a VAX 11/750, for systems of between 300 and 12000 atoms. Parallelism was also applied to crystallographic problems. Parallel versions of PROLSQ, a least-squares crystallographic refinement program, and BRUTE, a translation function program for molecular replacement, were written for a Connection Machine 200. In both cases a significant improvement in performance was produced.

The implications of the work were considered. It was possible to make some limited conclusions about the molecular mechanism of male rat hydrocarbon nephropathy. The risk to the human kidney from hydrocarbons was assessed using both biochemical and structural data. The problems encountered in crystallographic and molecular modelling studies were discussed, and possible improvements suggested. Finally, the benefits to be gained from parallel processing, with particular respect to the study of molecular structure, were presented.

Table of Contents

List of Figures

List of Tables

Abbreviations

Proteins

a2u	: Rat α_{2u}-globulin
MUP	: Mouse major urinary protein
RBP	: Human retinol binding protein
BBP	: Bilin binding protein (from *Pieris brassicae*)
INSEC	: Insecticyanin (from *Manduca sexta*)
BLG	: β-Lactoglobulin
A1GP	: α_1-acid glycoprotein
A1MG	: α_1-microglobulin
APH	: Aphrodisin
A2UREL	: α_{2u}-globulin related protein
PURP	: Purpurin
EPID	: Rat epidiymal secretory protein
PP14	: Pregnancy protein 14
C8γ	: Complement protein C8γ
PBP	: Bovine pyrazine binding protein
OBP	: Rat odorant binding protein
BG	: Bowman's gland protein
VEG	: Von Ebner's gland protein
PGDS	: Prostaglandin D synthetase
apoD	: Apolipoprotein D
Ch21	: Chondrocyte 21 protein
PBSN	: Probasin
FABP	: Fatty acid binding protein
cRBP	: Cellular retinol binding protein
cRABP	: Cellular retinoic acid binding protein

STVN : Streptavidin

PHY : Photoactive yellow protein

BSA : Bovine serum albumin

FLG : Feline β-lactoglobulin

1SA : Model 1 after simulated annealing

ADH : *Drosophila* alcohol dehydrogenase

HSD : 3α,20β-hydroxysteroid dehydrogenase

LT : *E. Coli* Heat labile enterotoxin

PT : *Bordatella pertussis* toxin

Small Molecules

TMP : 2,2,4-trimethylpentane

THBS : 3,5,5-trimethylhexanoyloxybenzene sulphonate

244T2 : 2,4,4-trimethylpentan-2-ol

244T1 : 2,4,4-trimethylpentan-1-ol

224T1 : 2,2,4-trimethylpentan-1-ol

DLO : *d*-limonene-1,2-oxide

DL : *d*-limonene

5HT : 5-Hydroxytryptamine

NAN-190 : 1-(2-methoxyphenyl)-4-[4-(2-phthalimido)butyl]-piperazine.HBr

NAN-190OH : 1-(2-hydroxyl)-4-[4-(2-phthalimido)butyl]-piperazine.HBr

NAN-190I : 1-(2-iodo)-4-[4-(2-phthalimido)butyl]piperazine.HBr

NAN-190E : 1-(2-methoxyphenyl)-4-[4-(2-phthalimido)but-2E-enyl]-piperazine.HBr

NAN-190Y : 1-(2-methoxyphenyl)-4-[4-(2-phthalimido)but-2E-ynyl]-piperazine.HBr

IMD-1 : 1-(2-methoxyphenyl)-4-[4-(4-I-benzamido)butyl]-piperazine.HBr

8-OH-DPAT : 8-Hydroxy-2-(di-*n*-propylamine)tetralin

Symbols

K_i : Inhibitory binding constant

K_a : Association constant

K_m : Dissociation constant

M : Molarity

Å : Angstrom

fs : femto second

K : Kelvin (temperature) or computing kilo (1024 elements)

rms : Root mean square

M : Molar

E : Energy

r : Distance

ϵ_0 : Dialectric Constant

ϵ_r : Relative Dialectric Constant

μ : Dipole

ΔG : Change in Gibbs free energy

ΔH : Enthalpic Energy

ΔS : Entropic Energy

T : Temperature

K_b : Force constant for bond terms

K_θ : Force constant for angle terms

K_η : Force constant for proper dihedral terms

K_ϕ : Force constant for improper dihedral terms

q : Charge

V : Potential

$\mathbf{x}$: Atomic coordinates

F : Force

λ_k : Scalar descent step size in steepest descents minimisation

$\mathbf{s}_k$: Descent direction in steepest descents minimisation

g : Gradient

b_k : Weighting factor used in conjugate gradients minimisation

δt : Time step in molecular dynamics simulations

$\mathbf{x}$" : Instantaneous acceleration

$\mathbf{v}$: Average velocity

$\mathbf{x}$' : Instantaneous velocity

$\mathbf{a}$: Average acceleration

$\mathbf{r}$: Cartesian coordinates

d : Correction applied to satisfy constraint in SHAKE

hkl : Miller Indices in X-ray Diffraction

F_{hkl} : Complex Structure Factor in X-ray Diffraction

xyz : Fractional Atomic Coordinates

U_{eq} : Isotropic Temperature Factor

R : Residual (crystallographic or otherwise)

B : Isotropic temperature factor

F_o : Observed structure factor

F_c : Calculated structure factor

$\langle F_c \rangle$: Time averaged structure factor

f : Atomic scattering factor

$\mathbf{h}$: Reflection hkl

λ : Wavelength of X-rays

θ : Angle of diffraction

G : Molecular scattering factor

R_j : Rotation matrix for symmetry operator j

Δ : Translation in brute force translation search

Miscellaneous

API	: American Petroleum Institute
SDS	: Sodium dodecyl sulphate
PAGE	: Polyacrylamide Gel Electrophoresis
PBS	: Phosphate Buffered Saline
AS	: Ammonium Sulphate
PEG	: Polyethylene glycol
PDB	: Brookhaven Protein Data Bank
NMR	: High-Field Nuclear Magnetic Resonance
FFT	: Fast Fourier Transform
SA	: Simulated Annealing
HiPPI	: High performance parallel interface
SPC	: Single point charge model
M_r	: Relative molecular mass
SISD	: Single instruction single data
SIMD	: Single instruction multiple data
MISD	: Multiple instruction single data
MIMD	: Multiple instruction multiple data
CS	: Meiko Computing Surface
ECS	: Edinburgh Computing Surface
CM-200	: Thinking Machines Corporation Connection Machine 200
DAP	: Active Memory Technology Distributed Array Processor
T800	: Inmos T800 Transputer
ALU	: Arithmetic logic unit
FPA	: Floating point accelerator
PE	: Processing element

Chapter 1

Introduction

Damage to the environment is an issue of constantly increasing importance. Much media attention has focused on the more visual effects, notably deforestation, acid rain, pollution, and endangered species. However, there are less obvious effects which may present long term problems for man. In particular the effects of exposure to chemicals and radiation must be an area for some concern. Man is constantly exposed to both natural and synthetic compounds which may be harmful. Understanding mechanisms of toxicity and identifying those biological systems which are in danger is a difficult task. Work from many different disciplines is required to obtain a clear picture of the toxicological effects of any particular compound or new combinations of compounds.

The background to this thesis arises from a study of kidney damage in male rats induced by small hydrocarbons. This kidney damage (nephropathy) is often associated with increased levels of renal carcinomas. The hydrocarbons are commonly found in petroleum based products. This link caused concern that petroleum compounds could be responsible for some of the renal cancers observed in man. The kidney performs many vital functions and its functioning is therefore decribed in some detail in the following chapter. The biochemical and toxicological background to male rat nephropathy is also covered. One protein is suggested to be primarily responsible for the effects seen in male rats, α-2u-globulin (a2u). This protein is considered in relation to other members of the super-family to which it belongs (the lipocalycins).

The mechanism through which this nephropathy is thought to occur, depends on an increase in the stability of a2u to lysosomal protease degradation upon hydrophobic ligand binding. Understanding how this binding can affect protein stability in this case requires examination of the structure of a2u both with and without bound ligands. Attempts to determine the three-dimensional structure of a2u by X-ray crystallography are presented. Problems with data collection associated with crystal twinning led to studies of a closely related protein from mice, mouse urinary protein (MUP). A brief background to X-ray crystallography is given to acquaint the reader with the technique. The determination of two small molecule structures is given to illustrate the technique. The diffraction data for both molecules is available from the author on request. The practical work involved in purification of MUP, and crystallisation of a2u are presented.

The modelling of protein structures using information from existing structures is of increasing interest. This is because the technology is now available and sequences are available for many proteins with interesting biological properties. Modelling their structures provides a 'short-cut' to structure determination by X-ray crystallography. A background to molecular modelling, and some of the techniques commonly used is given. This is followed by the modelling of the structure of a2u from the structures of related proteins. Deciding whether a model structure is close to the native structure is difficult. Various methods were used in the case of a2u, and these are assessed in some detail. Validation of the structure is only completely possible by comparison with the crystallographic structure. The coordinates for the related protein, MUP, became available allowing an assessment of the model structures to be made.

Both the modelling and crystallographic determination of structures require significant computing power. This became apparent during the modelling work and led to the investigation of parallel computing architectures. A background to parallel processing with specific machine examples is given. A parallel implementation of a standard energy minimisation and molecular dynamics code was written for a parallel computer. The work involved in writing this code, and the results obtained are presented. The application of data parallel computing to

two crystallographic problems is also investigated. All code developed during the course of this thesis is available directly from the author.

Finally, the discussion returns to the problems of kidney damage in both man and rats. The danger to the human kidney from hydrocarbons is assessed in the light of biochemical evidence, and the structural information available. The problems encountered in both crystallographic work and molecular modelling are discussed. The possible involvement of other members of the lipocalycin family in pathological states are considered. The advantages to be gained from parallel computing in structural studies are presented, with some possible future applications.

Chapter 2

The Kidney

2.1 The Biology of the Kidney

The kidney is responsible for a wide range of physiological functions: excretion, acid-base balance, extracellular fluid volume and osmolality homeostasis, and essential salt regulation. It is most commonly considered in terms of its excretory role which removes metabolic waste products and foreign substances via the urine. Although this is of major importance its other homeostatic roles are of vital significance. Damage to the kidney which impairs its function in any way will have serious repercussions for the whole body. The kidney is particularly susceptible to the effects of toxic agents in the circulation. This is primarily because of the large amount of blood which flows through the kidney (in the human about 1300 ml per minute). Obviously this high blood flow is necessary for kidney function; the purification of the blood. In addition certain regions of the kidney are involved in active metabolic transport which can result in the specific build up of toxins in some kidney cells. It is also known that the kidney has some of the same detoxification systems found in to the liver. These are capable of metabolically activating some chemical agents, converting them from nontoxic to highly reactive and toxic forms. The vulnerability of the kidney to toxic injury and its central role in whole body homeostasis makes the issue of toxic chemical mediated kidney damage (nephrotoxicity) vitally important. What follows is a brief background to the kidney with particular attention to the rat.

2.1.1 Gross Anatomy

In mammals there are two kidneys, generally bean-shaped, that lie behind the peritoneum at the rear of the abdomen on each side of the vertebral column. Located on the concave surface of each kidney is a slit called the hilus, through which passes the renal artery and vein, the lymphatics, a nerve plexus, and the renal pelvis (figure 2–1). A longitudinal section of the kidney shows two distinct regions; a dark outer region (the cortex), and a pale inner region (the medulla). The medulla is divided into striated conical masses called renal pyramids. The base of each pyramid is positioned on the corticomedullary boundary. The apex of each pyramid is called a papilla which extends towards and projects into the renal pelvis. It is from this pelvis that urine flows into the bladder. There are species differences in the number of renal pyramids: humans have between 8 and 18 (multilobar), while rats have only one (unilobar). The renal medulla can be divided into an outer and inner region. The outer medulla can be further divided into an outer and inner stripe. These stripes are particularly prominent in the unilobar kidney of the rat.

2.1.2 Microanatomy

The basic structural unit of the kidney can be considered to be the uriniferous tubule (figure 2–2). It is composed of a long, highly complex tubular structure called the nephron and a system of collecting ducts. Both of these can be further divided into several distinct morphological segments. Each nephron begins with a renal corpuscle (often referred to as a glomerulus). This is followed by: a proximal convoluted tubule; a proximal straight tubule; a thin limb; a thick ascending limb; and a distal convoluted tubule. The nephron is joined to the collecting duct by the connecting tubule. The collecting duct can be divided into cortical, medullary and papillary regions. There are approximately one million of these uriniferous tubules in each human kidney, as compared to about 30,000 in each adult rat kidney.

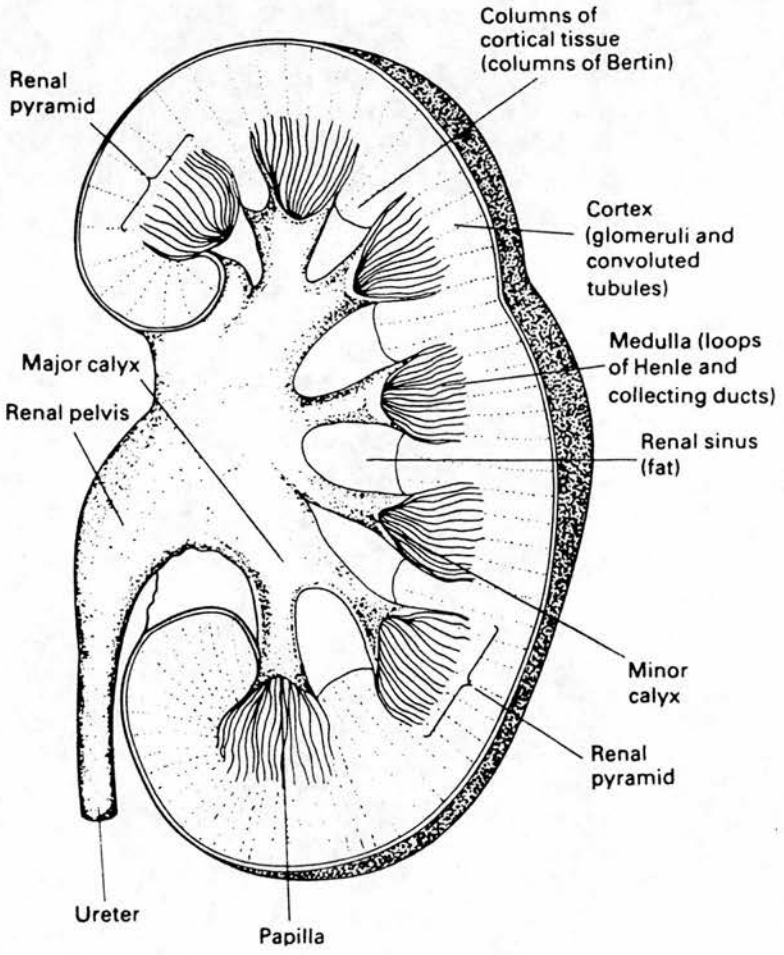


Figure 2–1: The gross anatomy of the human kidney (from Sweny *et al.*, 1989).

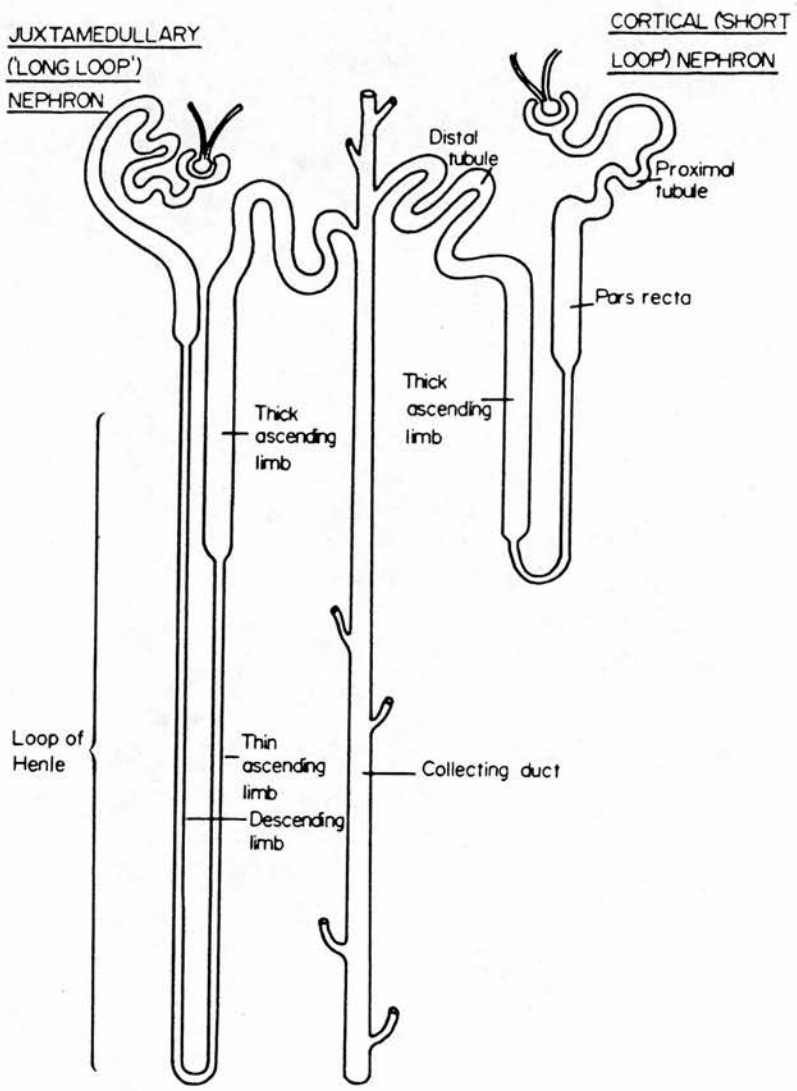


Figure 2–2: Diagrammatic representation of the uriniferous tubule (from Fourman and Moffat, 1971).

The Renal Corpuscle

This is composed of a bed of capillaries surrounded by a double-walled capsule known as Bowman's capsule. The inner layer of the capsule intricately covers the glomerular capillaries, generally separated from them only by the glomerular basement membrane. The space between the inner and outer layers of the Bowman's capsule is known as the Bowman's space. This space is continuous with the proximal tubule.

The Proximal Tubule

The proximal tubule is most frequently divided into a convoluted region, found in the cortex, and a straight region, found in the medulla. Evidence obtained from a variety of techniques including light, phase contrast, and fluorescence microscopy, as well as histochemistry and electron microscopy indicate that the proximal tubule of the rat can be divided into three distinct morphological segments. These are P_1, P_2, and P_3 as described in the rat by Jacobsen and Jørgensen (Jacobsen and Jørgensen, 1973). The following therefore relates to the rat as this is by far the best studied system.

The P_1 Segment

This makes up the initial portion of the convoluted tubule. The cells are generally columnar to cuboidal in shape. The brush border is well developed (2-3 μm) in the rat and less so in man and mouse (1 μm). The endocytotic apparatus is well developed and apical vacuoles are large and numerous. The Golgi apparatus is well developed and elements of the smooth and rough endoplasmic reticulum are present. Lysosomes in the rat range from small and dark (1 μm) to giant and light (5-7 μm), the latter only being seen in the female. In the human many more smaller (2 μm or less) lysosomes are seen. Mitochondria are numerous and elongated in all 3 species, lying perpendicular to the basement membrane. Peroxisomes are numerous in the human but are fewer and smaller in the rat and mouse.

The P_2 Segment

This makes up the remainder of the convoluted tubule and the initial portion of the straight tubule. A number of subtle features distinguish the P_2 and P_1 cells. Both cells and brush border are shorter in P_2 than P_1. There is still a well developed endocytotic apparatus though there are fewer apical vacuoles in the rat. The Golgi complex is well developed and is often more conspicuous in P_2 than P_1 in the rat. The lysosomes are relatively large and dark staining, in the rat, although no gaint lysosomes are seen. Mitochondria are typically fewer and shorter in the rat. Peroxisomes are more numerous in P_2 than P_1 in the rat and at least as numerous in the human.

The P_3 Segment

This makes up the remainder of the proximal straight tubule. The cells are lower in height and much less elaborate in shape than the cells of either P_1 or P_2. In the rat there is a dramatic increase in the height of the brush border in this segment, but in the human it becomes shorter. The endocytotic apparatus is poorly developed in all 3 species and mitochondria are shorter and more randomly oriented. In the rat the Golgi complex and endoplasmic reticulum are best developed in this segment but are less obvious in humans. In both man and rats the lysosomes are fewer and smaller. In the human there are fewer peroxisomes in P_3 but in the rat peroxisomes are larger and more numerous in this segment. These histological differences are supported by histochemical staining studies with specific markers for lysosomes and mitochondria. The general changes, in the rat, travelling along the proximal tubule from glomerulus to Loop of Henle: there is a reduction in cell height, an increase in the length of the brush border, and a reduction in the number and size of apical vacuoles, lysosomes and mitochondria.

The Thin Limb of the Loop of Henle

The structure of the thin limb, in both rats and humans, depends on the location of the renal corpuscle from which the nephron derives. Nephrons originating near the medulla have long loops of Henle. The thin limb descends into the inner medulla, follows a hairpin turn, then ascends back towards the cortex.

The Ascending Thick Limb

This lies between the end of the Loop of Henle and the renal corpuscle from which the nephron originated. The ascending thick limb becomes closely associated with afferent and efferent arterioles of the glomerulus, forming the juxtaglomerular apparatus. The cells at this point are narrow with their nuclei close together producing a dense staining with hematoxylin. This region is therefore referred to as the macula densa. The ascending thick limb continues from the macula densa for variable distance before it becomes the distal convoluted tubule.

The Distal Convoluted Tubule

This has a highly complex course within the cortex. As in the P_1 and P_2 regions numerous mitochondria are seen orientated perpendicular to the basement membrane. Distal convoluted tubules join to form the connecting tubule, which empty directly into the cortical collecting ducts.

The Collecting Duct

This can be divided into a cortical, medullary, and papillary portion. A single collecting duct drains a large number of nephrons. As the ducts descend through the medulla they join together forming larger and larger tubules. The larger papillary collecting ducts (ducts of Bellini) empty into the renal pelvis, which is continuous with the extra-renal collecting system.

2.1.3 Renal Function

The formation of urine in the mammalian kidney is the result of three processes: filtration, reabsorption, and secretion. Filtration occurs in the renal corpuscle. Each day approximately 180 litres of ultrafiltrate are formed by the human kidney yet only 1-2 litres of urine are excreted per day. The kidney therefore must reabsorb nearly 99% of the ultrafiltrate it produces. This is not just the reabsorption of water: many different salts and proteins also need to be taken back from the ultrafiltrate. The process of ultrafiltration is very indiscriminate. Molecules of less than 70,000 daltons pass into the ultrafiltrate. Many of these molecules are waste products to be removed from the body, but also included are substances vital to life. The uriniferous tubule therefore functions to reabsorb selectively those substances of importance. In addition tubular secretion helps eliminate compounds faster than by filtration alone. The role of the kidney in osmotic and extracellular fluid volume homeostasis is achieved by the excretion and reabsorption of specific substances under tight regulation.

The Renal Corpuscle

This is the site of ultrafiltrate formation from the plasma passing through the glomerular capillaries. These high pressure capillary beds result in the filtration of approximately 20% of the glomerular plasma across the glomerular filtration barrier into the Bowman's space. The barrier, composed of the fenestrated endothelium, the glomerular basement membrane, and the visceral layer of Bowman's capsule, limits the passage of substances on the basis of size, shape and charge. The glomerular filtration rate may be controlled, in part, by changes in the contractile state of the mesangial cells. These cells in addition with macrophages and resident immune response cells may help clear particulate and large molecular weight complexes from the glomerular basement membrane.

Proximal Tubule

The proximal tubule is responsible for reabsorbing the bulk, as much as 80%, of the produced ultrafiltrate. The P_1 segment rapidly changes the composition of the ultrafiltrate. The reabsorption of fluid is driven by active sodium ion transport. This creates an osmotic gradient which results in the passive movement of water out of the tubule lumen and eventually into peritubular capillaries. Changes are isosmotic with respect to the sodium concentration in the tubule, but bicarbonate, glucose, lactate and amino acid levels decrease rapidly within a few millimeters of the renal corpuscle. In the P_2 segment the same active transport of sodium and other small molecules occurs but with a much higher affinity for the latter due to the decreased concentration of these substances. The well developed lysosomal-phagosomal system in this section of the proximal tubule is important for the uptake of protein from the tubular fluid. The sex related differences in protein production, and hence filtration, in the rat are mirrored by differences in the lysosomal system between male and female. In the rabbit the P_2 section appears to be the site of organic acid secretion into the tubular fluid. The reabsorption of sodium chloride and other solutes continues in the P_3 segment. There is also passive secretion of urea which may be involved in urea recycling to deeper medullary structures.

The Loop of Henle

This structure is involved in the concentration of urine. On the downward section of the loop water passes out of the tubule due to the increasingly hypertonic surroundings. The permeability to sodium and urea is low, this results in the tubular fluid becoming hypertonic to plasma. On the ascending section of the limb the permeability to water is low but to sodium is high. Therefore sodium, followed passively by chloride, leaves the tubular fluid. The urine therefore becomes more dilute, and marginally more concentrated with respect to urea.

The Ascending Thick Limb

The tubular fluid is further diluted by active transport of sodium chloride. This transport of solutes into the interstitium contributes to the hypertonicity in the regions surrounding the loop of Henle.

The Distal Convoluted Tubule

This region plays an important role in the regulation of the sodium concetration in the extracellular fluid and acid-base balance. The sodium concentration is under the hormonal regulation of aldosterone, which stimulates increased sodium reabsorption by the distal convoluted tubule.

The Collecting Duct

The collecting duct is primarily involved in modulating the volume and osmolality of the extracellular fluid. This is accomplished through response to antidiuretic hormone. This hormone modulates the water permeability of the duct.

2.1.4 Sex Related Differences in Rat Kidney

The most significant difference between male and female rat kidney is related to urinary protein excretion. In young male rats the majority of the urinary protein is represented by a sex-dependent protein, alpha-2-urinary globulin. This protein is produced mainly in the liver under hormonal regulation by testosterone. Due to its small size it passes easily through the glomerular filttration barrier into the ultrafiltrate. Approximately 60% of the filtered protein is reabsorbed by the proximal tubule. After 6 months the urinary secretion of albumin begins to exceed that of a2u. This appears to be related to glomerular nephrosis or old rat nephropathy, an atrophy of the renal corpuscle and proximal tubule, in older male rats. This nephrosis probably causes an increase in glomerular permeability

thus allowing a larger amount of plasma albumin to be filtered into the ultrafiltrate. This nephrosis is seen to be delayed and less severe in female rats. The proximal tubule shows differences in ultra-structure probably because of the high level of urinary protein secretion in male rats. In the P_1 and P_2 regions apical vacuoles are larger and more numerous in the male rat. In the P_2 region the number and total volume of lysosomes is higher in male rats. The male rat has more of the cellular machinery required to take up protein from the ultrafiltrate and process it for return to the blood stream. There is also a larger volume of peroxisomes in the P_3 region in male rats.

2.1.5 Renal Metabolism of Endogenous and Exogenous Compounds

The kidney has at least three systems for the biotransformation of both endogenous and exogenous compounds. Their function is presumably to help in the detoxification of potentially harmful compounds. However, it is clear that in some cases the transformation can activate certain compounds from a relatively harmless to a toxic state.

Mixed Function Oxidase System

This is a non-specific microsomal enzyme system capable of hydroxylating a wide variety of hydrophobic molecules. This oxidation to more polar metabolites prepares the molecules for further reactions and/or excretion. The oxidase system is based around a renal cytochrome P_{450} similar to that found at higher levels in the liver. The enzymes involved are localised to the P_3 section of the proximal tubule. There are differences between the substrate specificity and enzyme activity of the renal and liver cytochrome P_{450} systems. The renal enzyme has generally lower activity except for a high capacity for ω and ω-1 hydroxylation of medium chain fatty acids.

Conjugation Reactions

These reactions result in highly polar metabolites which can be easily excreted from the body. The kidney has the capacity for the formation of glucuronide, sulphate, and glutathione conjugates.

Prostaglandin Endoperoxide Synthetase Co-oxidation

This type of co-oxidation is the result of a radical forming process which produces activated intermediates capable of reacting with DNA, RNA and proteins, resulting in cell injury and death. This may be an important mechanism in the action of certain nephrotoxic agents, for example acetaminophen.

2.1.6 Relevance of the Rat Kidney to Human Kidney

The old rat nephropathy seen in male rats, and eventually in female rats, has a resemblance to human arteriolar nephrosclerosis and chronic pyelonephritis. As in the rat, the human kidney exhibits a decreasing glomerular filtration rate with age, diminished capacity to conserve sodium and a diminished ability to concentrate urine. The human glomeruli show sclerotic changes coupled with tubular atrophy and interstitial inflammation. However, the principle lesions in the human appear to involve the blood vessels rather than the renal corpuscle and tubule in rats. Old rat nephropathy may have a part to play in renal carcinogenesis. The increased mitotic rates and chronic inflammation in nephropathic regions may well act as promoters of carcinomas. From this it is clear that any damage to the kidney may result in an increased incidence of renal carcinomas. It is reported that there is a greatly increased adenocarcinoma incidence in human dialysis patients.

2.2 Hydrocarbon Induced Nephropathy in Rats

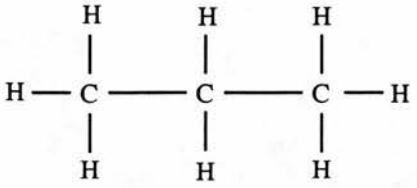
A wide variety of chemicals are known to cause kidney damage in rats ranging from cell necrosis to carcinomas. In many cases this kidney damage is eventually fatal to the rat. It is probable that many of these chemicals directly promote renal neoplasia, tris(2,3-dibromopropyl)phosphate, aflatoxin B_1, nickel subsulphide, dimethylnitrosamine, endrin, dieldrin, to name only a few. In contrast there are chemicals which are seen to produce a nephropathy in specific regions of the male rat kidney. This nephropathy can be associated with an increased incidence of renal carcinomas. The chemicals which produce these effects are usually small and hydrophobic and often derived from petroleum.

2.2.1 History

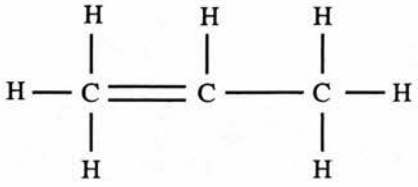
In the mid 1970's the American Petroleum Institute (API) initiated studies into the effect of short term exposure to petrol vapours on various animals. The results appeared to be essentially negative. In 1978 a long-term exposure study was started on mice and rats. The results showed statistically significant evidence of kidney cancer in those male rats exposed to vapourized unleaded petrol for 6 hours a day, 5 days a week, over a 27 month period. In addition, dose related levels of kidney damage were seen in all male rats. The obvious question was "Does normal exposure to petrol vapour represent a significant human health hazard?".

2.2.2 Petroleum

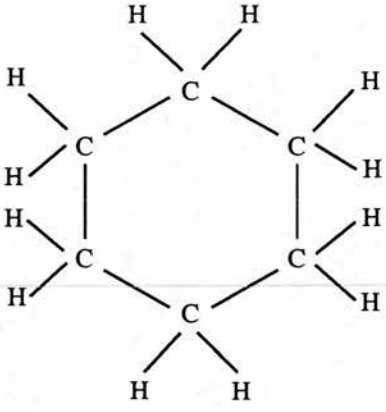
The majority of people in the developed Western world rely on petroleum products for the convenience and comfort of their lifestyle. Petroleum products are ultimately derived from crude oils, which are composed of thousands of chemical compounds which are mainly hydrocarbon based. There are 4 main hydrocarbon groups: alkanes, alkenes, naphthenes and aromatics, examples of



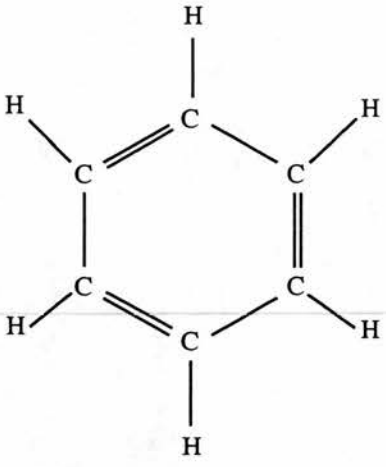
Propane



Propene



Cyclohexane



Benzene

Figure 2–3: Examples of the four main types of hydrocarbon found in unleaded petrol

which are shown in figure 2–3. The wide variety of hydrocarbon compounds gives, theoretically, at least 1500 different components in petrol. The API PS-6 vapourized petrol inhalation study (MacFarland, 1983) identified 151 components, 42 of which account for 75% of the petrol volume. Petrol is derived from the blending of several different crude oil refinery streams. The mixture produced is designed to meet certain performance characteristics in modern cars, such as octane level, volatility, and complete combustion. The composition of petrol therefore varies from different producers but has many components in common.

2.2.3 The PS-6 90-day and 2-year Petrol Studies

The API study into the effects of hydrocarbons comprised both short term and long term exposure experiments.

The 90-day Study

Sprague-Dawley rats and squirrel monkeys were exposed to both vapourised unleaded and leaded petrol for 6 hours per day, 5 days per week for 13 weeks. Various tests of response were used, such as organ weight, pulmonary function, changes in the blood, and the histopathological examination of over 30 tissues. Some sex dependent and species dependent differences were seen for some of the tests but no changes which were thought to be toxicologically important. However, subsequent re-examination of kidney tissue samples showed a discernible change in the male rats exposed to high levels of unleaded petrol. These changes consisted of increased regeneration of some cells and dilated tubules containing proteinaceous material.

The 2-year Study

Fischer 344 rats and B6C3F_1 mice were exposed to different levels of vapourised unleaded petrol for 6 hours per day, 5 days per week, for between 24 and 26 months. Animals were sacrificed at 3, 6, 12, and 18 months and at the termination of the study. At the 3-month and 6-month points a progressive increased incidence of renal disease was seen with tubular degeneration and regeneration coupled with tubular dilatation. At the 12-month point, the start of old rat nephropathy, a progressive glomerulonephrosis became evident. These changes obscured the earlier changes seen at 3 and 6 months. At termination nearly all male rats and many female rats showed advanced glomerulonephrosis. In addition, at termination, a dose related occurrence of primary renal neoplasms was seen in the male rats.

2.2.4 Further Studies

The PS-6 study indicated that long term inhalation of petrol by male rats led to kidney damage and an increase in kidney cancers in a dose related way. These results prompted further research by other groups. Their results showed similar tubular degeneration and regeneration, tubular dilatation and angular protein (hyaline) droplets in the cytoplasm of proximal convoluted tubule cells. The term hyaline droplet is a general description of small spherical structures that are visible with hematoxylin and eosin staining within the cytoplasm of renal tubular cells (Oliver and MacDowell, 1954, 1958). These were produced on exposure to decalin (Gaworski *et al.*, 1982), JP5 jet fuel (MacNaughton and Uddin, 1983), and many commercially available light hydrocarbon mixtures (Phillips and Cockrell, 1983). A study by the Standard Oil Company using different naphtha fractions showed that similar results were obtained with high alkane content fractions. The renal lesions observed typically consisted of groups of dilated tubules located at the cortico-medullary junction. The dilation was caused by accumulation of granular proteinaceous material, probably cellular debris, in the lumens of these tubules. Seen concurrently were scattered groups of degenerating and regenerating proximal convoluted tubules in the renal cortex. Hyaline droplets were observed in the epithelial lining cells of the proximal convoluted tubules, of male rats only. It was first thought that the hyaline droplet formation was a by-product of tubular degeneration and regeneration, but subsequent studies showed the latter without droplet formation, suggesting a different mechanism for their production. A study by the Exxon Corporation with C_{10}-C_{11} isoparaffin showed the hyaline droplet accumulation in proximal renal tubules seen previously (Phillips and Cockrell, 1983). The droplets occurred with increasing severity in a dose related manner, starting previous to 5-day sacrifice. Electron microsopy of sections of affected proximal renal tubules showed electron dense phagolysosomes. These phagolysosomes corresponded to the hyaline droplets as identified by toluidine blue staining of the same cells. Studies with the volatile hydrocarbon decalin showed a similar increase in hyaline droplet formation (Alden *et al.*, 1983). The proteinaceous material present in the

droplets was identified as alpha-2-urinary globulin (a2u) by two-dimensional electrophoresis followed by Western Blotting, and immunofluorescence of tissue sections. These results led to the conclusion that the inhalation of hydrocarbons leads to a change in the normal processing of a2u reabsorbed in the kidney of male rats. In fact spontaneous hyaline droplets are seen in the kidneys of healthy male rats, probably due to an inefficiency of protein reabsorption/catabolism (Logothetopoulus and Weinbren, 1955). However, the results did not show how the hydrocarbon compounds caused this change in a2u processing, several possibilities were put forward (Alden et al., 1983):

- Hydrocarbon activation of a2u production at the hepatocyte.
- Formation of a undigestible hydrocarbon (or its metabolites) protein complex.
- Hydrocarbon inhibition of proximal convoluted tubule lysosomal activity.
- A combined mechanism involving the above three.

2.2.5 Studies of Hyaline Droplet Nephropathy *in vivo*

Studies of several components of unleaded petrol demonstrated that branched aliphatic alkanes are primarily responsible for the nephrotoxic action of unleaded petrol (Halder *et al.*, 1983). The common component of petrol, used as the reference compound for octane rating, 2,2,4-trimethylpentane (TMP) was seen to be very active as a nephrotoxic agent (Halder *et al.*, 1985). This compound has therefore been used as a prototype compound to study the mechanism of unleaded petrol induced nephropathy (Short *et al.*, 1986; Stonard *et al.*, 1986; Loury *et al.*, 1987). It was shown that 72 hours after a single oral dose of radiolabelled [^{14}C]TMP, more radiolabelled material was present in the kidneys of male than female rats (Kloss *et al.*, 1985). There has also shown to be a parallel increase in the renal concentration of a2u as early as 4 hours following an oral dose of TMP (Stonard *et al.*, 1986). The work of Charbonneau *et al*

(Charbonneau et al., 1987) determined the fate of TMP in male and female rats, in terms of metabolic products and tissue localisation. Aliphatic hydrocarbons are oxidised to alcohols by the cytochrome P_{450} monooxygenase system (Jakoby *et al.*, 1982). The resulting alcohols can be oxidised to aldehydes by alcohol dehydrogenases and the further oxidised to acids by aldehyde dehydrogenases. Acid metabolites with long or complex aliphatic chains may be hydroxylated to form hydroxyacids. Analysis of urine from male rats treated with TMP shows a variety of metabolites: trimethyl-branched pentanols, pentanoic acids, and hydropentanoic acids (Olson *et al.*, 1985). They suggest that metabolites of TMP are produced by 3 different pathways, where TMP undergoes oxidation of carbon 1,4 or 5. Analysis of kidney, liver, and plasma tissues from male and female rats dosed with 2,2,4-trimethyl[5-^{14}C]pentane showed a male specific retention of TMP derivatives in the kidney. In addition the concentration of TMP derivatives in male kidney, liver and plasma remained constant for 12-24 hours after dosing, while they declined rapidly in female rats. Although, the final percentage of the initial dose excreted in the urine after 48 hours was the same in both male and female rats. The renal concentration of a2u in male rats was seen to increase in a dose related way after TMP treatment. Two metabolites, 2,4,4-trimethylpentan-2-ol (predominantly) and 2,4,4-trimethylpentanoic acid were detected in male rats kidneys but not females, although conjugated 2,4,4-trimethyl-2-pentanol was detected in female rat urine. The results suggest that no difference exists in the metabolism of TMP in male and female rats. The concomitant renal accumulation of 2,4,4-trimethylpentan-2-ol and a2u suggests an association between these two components. Further work by Lock *et al* (Lock *et al.*, 1987) showed a reversible binding between a radiolabelled metabolite of [^{3}H]TMP and a protein fraction containing a2u. Dialysis of the fraction against buffers of pH 4.5 to 8.5 did not disassociate the radiolabel from the fraction. However, dialysis against SDS led to a significant decrease in the binding of radiolabel to the fraction. Ethyl acetate extraction of the a2u containing fractions recovered the majority (92%) of the radiolabelled material. Gas chromatography-mass spectrometry of the sample identified the metabolite as 2,4,4-trimethylpentan-2-ol (244T2). Analysis of 2,4,4-trimethyl-[^{3}H]pentan-2-ol

levels over time show a specific retention in the kidney, at 72 hours over 40% of the radiolabelled material was present as 2,4,4-trimethylpentan-2-ol. Previous to this work it had been suggested that a stable Schiff-base product formed in the liver between an aldehyde metabolite of TMP and the ϵ-amino residue of a lysine (or lysines) in a2u (Gibson and Bus, 1987). However, sodium cyanoborohydride treatment of the radiolabelled fractions did not alter the reversible nature of the binding, as it should have done if a Schiff-base were involved. Further work with [^{14}C]1,4-dichlorobenzene (Charbonneau *et al.*, 1988a) and [^{14}C]isophorone (Strasser *et al.*, 1988) showed coelution of radiolabelled metabolites with a2u. Again reversible binding was demonstrated by equilibrium dialysis. The metabolites were identified, by gas chromatography-mass spectrometry, as 1,4-dichlorobenzene, 2,5-dichlorophenol and isophorone. Exposure to 3,5,5-trimethylhexanoyloxybenzene sulphonate (THBS) has been shown to cause hyaline droplet nephropathy (Lehman-McKeeman *et al.*, 1991). The major metabolite in this case was identified as a *cis* γ-lactone of trimethylhexanoic acid. It has been shown that presence of a2u is neccessary for the promotion of kidney tumours in the male rat by *d*-limonene (Dietrich and Swenberg, 1991a). The inbred NBR rat strain does not synthesize a2u, as a result of a tissue and gene specific regulatory defect (Chatterjee *et al.*, 1989). Male NBR rats do not develop hyaline droplet nephropathy when exposed to the active chemicals mentioned above (Dietrich and Swenberg, 1991b).

2.2.6 Binding Studies *in vitro*

An binding assay *in vitro* was developed to characterize more fully the binding of chemicals to a2u (Borghoff *et al.*, 1988). Firstly it was demonstrated that binding of [^{3}H]244T2 to kidney cytosol is due to the presence of a2u, since binding was not observed after immunoprecipitation of a2u. It was also observed that two proteins related to a2u; bovine β-lactoglobulin (BLG), and rat α_1-acid glycoprotein (A1GP), were able to bind [^{14}C]244T2 whereas the nonrelated proteins β_2-microglobulin and lysozyme did not. The binding of [^{3}H]244T2 to a2u from the kidney of control rats was similar as the binding to a2u from

TMP-treated rats (Kd 132 nM). Using a competitive binding assay (competing against [^{3}H]244T2) it was possible to calculate apparent inhibition constants for a variety of known nephropathic hydrocarbon metabolites and other similar shaped molecules. It is seen that the range of apparent binding affinities (as represented by the apparent inhibition constants) is large for known nephrotoxic agents. They range from 3.9×10^{-4} M for 2,5-dichlorophenol to 5.1×10^{-7} M for *d*-limonene oxide (DLO), even though their nephropathic potency is relatively the same (Borghoff *et al.*, 1991). Comparison of the chemicals which bound suggests that several structural factors are important in determining binding affinity: a lipophilic region, an electronegative atom, and the correct steric volume.

2.2.7 Protein Degradation Studies

Proteins differ in their susceptibilty to hydrolysis by proteases (Rogers *et al.*, 1986). It is suggested that the half-lives of proteins are determined by specific molecular determinants in the protein (Dice, 1987). It is observed that rapidly degraded normal proteins have regions rich in proline, glutamate, serine and threonine (PEST regions), suggesting that the amino acid sequence may have an important role in determining the rate of proteolysis. The hydrolysis of a2u was studied relative *in vitro* to other proteins and also in the presence of 244T2 (Charbonneau *et al.*, 1988b). [^{14}C]a2u can be obtained by purification from kidney cytosol prepared from rats treated with a mixture of ^{14}C-amino acids. The hydrolysis of a2u, BLG, lysozyme, and high and low molecular weight proteins from rat liver was studied using either a mixture of purified proteases or pure proteinase K. All proteins except a2u were digested rapidly. Proteinase K digestion of [^{14}C]a2u prepared from TMP treated rats (244T2 bound a2u) was seen to be different to a2u from rats not treated with TMP. After 48 hours hydrolysis there was a 30% reduction in hydrolysis of the 244T2 bound a2u, even though only 31$\pm$6% of the a2u had 244T2 bound. These results suggest the binding of 244T2 to a2u forms a complex highly resistant to proteolysis by proteinase K. This increase in protein half life could be responsible for the accumulation of protein in lysosomes of the P_2 segment of the proximal

convoluted tubule, where proteolytic activity appears to be very high. Further *in vitro* digestion studies have been carried out with purified lysosomal proteases (Lehman-McKeeman *et al.*, 1990). The use of selective protease inhibitors showed that the cysteine proteinases (cathepsins B, H and L) and the aspartic acid proteinase, cathepsin D, are important in lysosomal breakdown of a2u. The binding of *d*-limonene or 1,4-dichlorobenzene did not alter the degradation of a2u. Binding of *d*-limonene-1,2-oxide, isophorone, or 2,5-dichlorophenol decreased degradation by approximately by 30%. The metabolites themselves were seen to have no effect on the activity of the lysosomal proteinases, confirming that a2u is not acting as a transporter of proteinase inhibitors. The results suggest that the presence of an oxygen function is required for a metabolite to alter the lysosomal degradation of a2u and thus produce a nephropathy, although the small number of compounds tested does not make this result conclusive.

2.2.8 Proposed Mechanism of a2u Nephropathy in Male Rats

A mechanism for a2u nephropathy in male rats has been proposed (Swenberg *et al.*, 1989). This is shown diagramatically in figure 2–4 and can be summarized thus:

- Reversible binding of certain hydrocarbon metabolites to a2u forms a complex which is highly resistant to lysosomal proteinase digestion.
- This leads to accumulation of the complex in P_2 renal epithelial cells, causing lysosomal protein overload and individual cell necrosis.
- This is followed by cell regeneration which continues as long as a2u is produced and the rat is exposed to the hydrocarbon.
- The increased amount of cell proliferation acts as a tumor promotor by clonally expanding spontaneously initiated cells in the kidney.

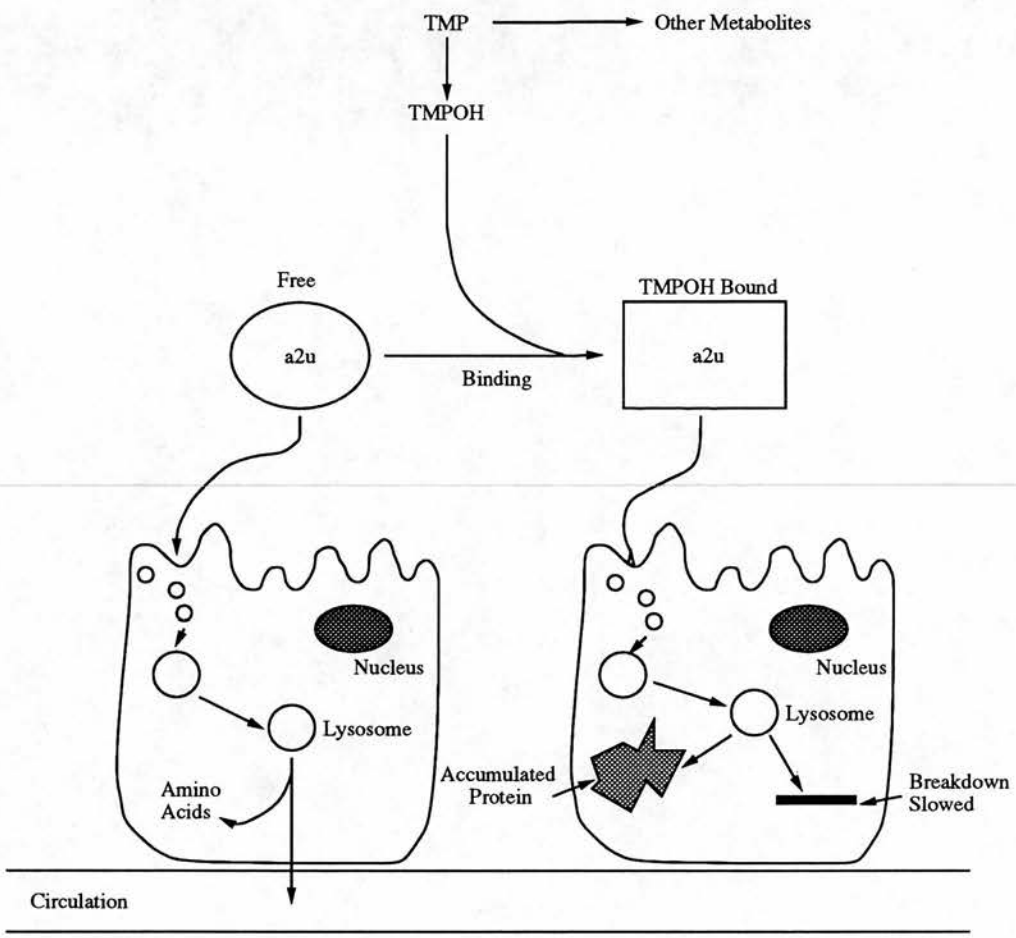


Figure 2–4: Proposed mechanism for alpha-2u-globulin hyaline droplet nephropathy induced by TMP in male rats.

2.3 Alpha-2-Urinary Globulin

Alpha-2-urinary globulin is a low molecular weight protein (18.5 kDa) that is synthesised in the parenchymal cells of the liver of adult male rats. The protein is secreted into the blood at a rate of 90-180 μg/g liver/hr (Antakly *et al.*, 1982). The synthesis of the protein is under a complex hormonal control. The male androgens are required for its synthesis: castration of the male decreases a2u production, and a2u production can be promoted in female rats by dosing with testosterone (Haars and Pitot, 1980). In addition synthesis is seen to be influenced by growth hormone, glucocorticoids, and thyroxine (Roy, 1973a, 1973b). Due to its low molecular weight, a2u is readily filtered through the glomerulus and into the ultrafiltrate. The kidneys of an adult male rat filter 50-60 mg of a2u per day (Neuhaus *et al.*, 1981). Of this 60% is reabsorbed and processed by the P_2 section of the proximal renal tubules. The remains 40% stays in the filtrate and eventually passes out in the urine. In contrast to the male, female rats excrete only 1% of the amount of a2u. In addition, the site of a2u synthesis is thought to be the salivary gland and not the liver in the female rat. The amino acid sequence of a2u has been determined by direct protein sequencing (Drickamer *et al.*, 1981) and also by the cloning and sequencing of cDNAs (Unterman *et al.*, 1981). The mature protein is seen to be 162 amino acids long (figure 2–5). The cDNA sequence shows a 19 amino acid leader sequence typical of proteins which are secreted extracellularly, this leader sequence is cleaved off prior to secretion into the blood stream. The a2u found in urine is seen to be the same as the 162 amino acid hepatic form. However, it is reported that the a2u found in the kidney is only 151 amino acids, some 11 residues shorter (Swenberg *et al.*, 1989). It is suggested that 9 amino acids are cleaved from the N-terminus and two amino acids are removed from the C-terminal part of the protein. The initial purification of a2u from male rat urine was by ammonium sulphate precipitation followed by DEAE cellulose chromatography. Analysis of the purified protein suggested a molecular weight of 26.4 kDa and an isoelectric point of 3.4 (Roy *et al.*, 1966). A certain amount of

```
-19                    1
  MKLLLLLLCL GLTLVCGHAE EASSTRGNLD VAKLNGDWFS IVVASNKREK

  IEENGSMRVF MQHIDVLENS LGFKFRIKEN GECRELYLVA YKTPEDGEYF

  VEYDGGNTFT ILKTDYDRYV MFHLINFKNG ETFQLMVLYG RTKDLSSDIK

  EKFAKLCEAH GITRDNIIDL TKTDRCLQAR G
                                    162
```

Figure 2–5: Amino acid sequence of rat alpha-2u-globulin. The mature protein sequence, from amino acid 1 onwards, is produced by cleavage of the proceeding leader sequence. The standard single letter amino acid codes are used.

carbohydrate was also detected in the purified protein fraction, suggesting that a2u was a glycoprotein. The purification method was subsequent refined to ammonium sulphate precipitation, gel filtration using Sephadex G-100, and then isoelectric focusing (Lane and Neuhaus, 1972). This improved methodology showed a2u to be non-glycosylated and of molecular weight between 20.0 and 21.0 kDa. In addition, several different charge species were detected varying in isoelectric point from pH 5.0 to 5.8, with the major component having an isoelectric point of 5.2. Antiserum prepared against this purified a2u was used to show that the male rat excreted approximately 20 mg of a2u per day in the urine, whilst the female excreted none.

2.4 Mouse Major Urinary Protein

Mouse urine, both male and female has been shown to contain a protein of the similar size to a2u (Finlayson *et al.*, 1968). The protein was purified from urine using gel filtration with Sephadex G-100, followed by ion exchange chromatography with DEAE. Analysis of the purified protein suggested a molecular weight of between 17.0 and 17.5 kDa. At least three charge species were identified by ion exchange chromatography. Tryptic peptide mapping of

these different forms showed two of them to be very similar, differing in only one peptide, whilst the third differed in several peptides. N-terminal sequencing of the same isoforms, as determined by agar gel electrophoresis, from inbred and wild mice showed identical sequences (Finlayson *et al.*, 1974). Analysis of a third isoform seen in the inbred mice showed a N-terminal sequence significantly different to the other two isoforms. Subsequent studies into the major mouse urinary allergen from mice identified a protein with molecular weight between 18.0 and 21.0 kDa (Lorusso *et al.*, 1986). This major allergen is thought to be identical to MUP on the basis of tissue localisation, size, charge heterogeneity, and sex related differences. Isoelectric focusing showed at least 4 charge species with the same antigenic activity, the major band had an isoelectric point of 3.9. Both the urine and serum of male mice was seen to have a four fold higher level of the antigen than female mice. The antigen was relatively resistant to both proteolysis and heating. Genetic analysis has shown that the the MUPs are encoded by a family of genes consisting of at least 35 to 40 highly homologous genes and pseudogenes (Bennett *et al.*, 1982; Bishop *et al.*, 1982). These are seen to be located in the Mup-a locus of chromosome 4, the genes for a2u are located on chromosome 5 which appears to be homologous to the mouse chromosome 4 on the basis of location of other genetic loci (Szpirer *et al.*, 1990). The nucleotide sequences of 5 mRNAs from different tissue sources have been determined; MUP I - MUP V (Shahan *et al.*, 1987a). The sequences suggest that the mRNAs are encoded by different members of the gene family. The amino acid sequences derived from the nucleotide sequences are 180 amino acids in length the first 18 residues being a signal peptide which is cleaved on secretion to produce a mature polypeptide of 162 amino acids (figure 2–6). The similarities and differences between the sequences suggest that MUP I and II can be classified as a group which are nearly identical (99.6%), whilst MUP III, IV, and V show greater difference between each other and the first group (figure 2–7). The tissue dependent expression of these mRNAs is complex in its distribution (Shanan *et al.*, 1987b). The biological significance of this distribution is unclear.
Comparison of the amino acid sequences of a2u and MUP shows a high level of sequence identity (figure 2–8). Despite this similarity in amino acid sequence and

gene distribution and structure, the biological similarities are less marked. The sex dependent differences seen in the rat are not seen in the mouse. Female mice do synthesise and excrete MUP, although at a 4 to 5 fold lower level than male mice. The difference in levels is thought to be due to the same kind of complex hormonal control seen in the rat (Hastie *et al.*, 1979). At first glance it would be expected that the hydrocarbon induced nephropathy seen in rats would also be observed in mice. However, no work to date has shown any nephropathy in mice, male or female, induced by the same or similar hydrocarbons to those toxic to the male rat. Oral dosing of both male and female mice with THBS did not induce hyaline droplet formation (Lehman-McKeenan *et al.*, 1991). The levels of THBS metabolites detected in the kidneys were much less than those seen in either male (50 times greater) or female (5 times greater) rats. The basis for the mouse resistance to hyaline droplet nephropathy has been studied using *d*-limonene as the hyaline droplet inducing agent (Lehman-McKeenan and Caudill, 1992). Analysis showed that rats and mice excrete similar levels of a2u (12.24±0.60 mg) and MUP (14.88±0.99 mg) daily in the urine. Both rats and mice metabolised the *d*-limonene to *d*-limonene-1,2-epoxide which is thought to be an active agent in rat hyaline droplet nephropathy. However, binding of this active metabolite to the proteins in mouse kidney was not detected, whereas about 40% of the metabolised *d*-limonene was bound to proteins of the male rat kidney. Binding studies *in vitro* with *d*-limonene-1,2-epoxide showed no binding to MUP. Perhaps the most important difference observed was in the renal handling of a2u and MUP. Rats reabsorbed about 60% of the a2u filtered in the kidney, whilst no renal reabsorbtion of MUP was seen in either male or female mice.

2.5 The Alpha-2-Urinary Globulin Superfamily

The similarity in amino acid sequence between a2u, α_1-acid glycoprotein (A1GP), plasma retinol binding protein (RBP), β-lactoglobulin (BLG), and α_1-microglobulin (A1MG) was noted some time ago (Pervais and Brew, 1987). This similarity of amino acid sequence was also reinforced by similarities in

```
-18                  1
  MKMLLLLCLG LTLVCVHAEE ASSTGRNFNV EKINGEWHTI ILASDKREKI

  EDNGNFRLFL EQIHVLENSL VLKFHTVRDE ECSELSMVAD KTEKAGEYSV

  TYDGFNTFTI PKTDYDNFLM AHLINEKDGE TFQLMGLYGR EPDLSSDIKE

  RFAQLCEEHG ILRENIIDLS NANRCLQARE
                                  162
```

Figure 2–6: Amino acid sequence of mouse major urinary protein. The mature protein sequence, from amino acid 1 onwards, is produced by cleavage of the proceeding leader sequence. The standard single letter amino acid codes are used.

intron/exon boundary distribution (Ali and Clark, 1988). At the time the crystallographic structures of both human plasma RBP and bovine BLG had been solved (Newcomer *et al.*, 1984; Papiz *et al.*, 1986). Although the sequence homology between these two proteins is low, the overall tertiary structure is very well conserved (Sawyer, 1987). Sequence alignments suggested that tobacco hornworm insecticyanin (INSEC), a biliverdin transport protein from insect haemolymph, was also a member of the same superfamily (Pervais and Brew, 1987). Crystallographic determination of the structure showed the same tertiary fold seen for RBP and BLG (Holden *et al.*, 1987). The structure of a homologous protein, bilin binding protein (BBP), was also determined by X-ray crystallography revealing the same fold (Huber *et al.*, 1987a,b). Recently mouse urinary protein (MUP) has been crystallised (Bocskei et al., 1991) and its three dimensional structure solved; revealing a lipocalycin tertiary fold. They also show the same overall topology as previous lipocalycin structures. On the basis of sequence alignments, tertiary structure and function, it has been suggested that the structural motif of the superfamily forms a basic framework for the binding and transporting of small lipophillic molecules (Godovac-Zimmermann, 1988). This family of proteins has therefore been called the lipocalycin, or lipocalin superfamily, because of the lipophillic nature of many of the ligands and the common teriary fold - a calyx (Sawyer and Richardson, 1990). This text

```
mupi   EEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQIHVLENSLVLKFHTVR
mupii  EEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQIHVLEKSLVLKFHTVR
mupiii ----------------------------------------------NSLVFKFHLIV
mupiv  EEATSKGQNLNVEKINGEWFSILLASDKREKIEEHGSMRVFVEHIHVLENSLAFKFHTVI
mupv   EEASSERQNFNVEKINGKWFSILLASDKREKIEEHGTMRVFVEHIDVLENSLAFKFHTVI
mupa   EEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQIHVLENSLVLKVHTVR
mupe   EEASSTGRNFNVEKINGEWHTIILAFDKREKIEDNGNFRLFLEQIHVLENSLVLKFHTVR
       ***.*. .*.*******.* .*.** *******..*..*.*.*.*.***.**..* * .

mupi   DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLY
mupii  DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLY
mupiii NEECTEMTAIGEQTEKAGIYYMNYDGFNTFSILKTDYDNYIMIHLINKKDGKTFQLMELY
mupiv  DGECSEIFLVADKTEKAGEYSVMYDGFNTFTILKTDYDNYIMFHLINEKDGKTFQLMELY
mupv   DEECTEIYLVADKTEKAGEYSVTYDGFNTFTILKTDYDNYIMFHLINKKDEENFQLMELF
mupa   DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLY
mupe   DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINENDGETFQLMGLY
       ..**.*.  ....***** * . *******.* ******..* ****..*...****.*.

mupi   GREPDLSSDIKERFAQLCEKHGILRENIIDLSNANRCLQARE
mupii  GREPDLSSDIKERFAKLCEEHGILRENIIDLSNANRCLQARE
mupiii GREPDLSLDIKEKFAKLCEEHGIIRENIIDLTNVNRCLEARE
mupiv  GRKADLNSDIKEKFVKLCEEHGIIKENIIDLTKTNRCLKARE
mupv   GREPDLSSDIKEKFAKLCEEHGIVRENIIDLSNANRCLQARE
mupa   GREPDLSSDIKERFAQLCEKHGILRENIIDLSNANRCLQARE
mupe   GREPDLSSDIKERFAQLCEKHGILRENIIDLSNANRCLQARE
       **..**. ****.*..***.***..******...****.***
```

Figure 2–7: Multiple sequence alignment of different mouse urinary protein amino acid sequences. Alignment was with the program CLUSTAL using default parameters. Identical matches across all sequences are indicated with an asterisk, conservative substitutions are indicated with a point. Sequences were obtained from the GenBank (release 70.0) and EMBL (release 29.0) sequence databases. Mupi (musmupi), mupii (musmupii), mupiii (musmup3b), mupiv (musmupiv), mupv (musmupv) are the sequences described by Shanan *et al.*, 1987b. Mupa (musmupa) is described by Kuhn *et al.*, 1984. Mupe (musmupe) is described by Bennett *et al.*, 1987.

```
                        1
a2u MKLLLLLLCLGLTLVCGHAEEASSTRGNLDVAKLNGDWFSIVVASNKREK
    ||: ||||||||||||.|||||||  |::|.|:||:| .|::||:||||
MUP MKM.LLLLCLGLTLVCVHAEEASSTGRNFNVEKINGEWHTIILASDKREK

a2u IEENGSMRVFMQHIDVLENSLGFKFRIKENGECRELYLVAYKTPEDGEYF
    ||:||.:|:|:::|.||||||.:||:.  ::||.|| :|| ||...|||
MUP IEDNGNFRLFLEQIHVLENSLVLKFHTVRDEECSELSMVADKTEKAGEYS

a2u VEYDGGNTFTILKTDYDRYVMFHLINFKNGETFQLMVLYGRTKDLSSDIK
    |.||| ||||||||||.::| |||| |:|||||||.||||..|||||||
MUP VTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLYGREPDLSSDIK

a2u EKFAKLCEAHGITRDNIIDLTKTDRCLQARG
    |:||.|||.||| |:||||||...:||||||:
MUP ERFAQLCEEHGILRENIIDLSNANRCLQARE
                                   162

          Gap Weight: 3.000       Average Match: 0.540
       Length Weight: 0.100    Average Mismatch: -0.396
             Quality: 200.6              Length: 181
               Ratio: 1.114                Gaps: 1
  Percent Similarity: 81.111   Percent Identity: 66.667
```

Figure 2–8: Pairwise alignment of the amino acid sequences of alpha-2u-globulin (a2u) and mouse major urinary protein (MUP). Alignment was with the UWGCG program GAP. Identical matches are indicated by a vertical line, conservative matches with a colon, and semi-conservative matches with a dot.

RBP	MUP	BLG	BBP	INSEC
70-174	64-157	66-160	9-119	9-119
120-129		106-119	43-175	43-175
4-160				

Table 2–1: Disulphide bonds seen in the lipocalycins by X-ray crystallography.

will use the more etymologically correct lipocalycin to describe members of this family of proteins.

2.5.1 The Lipocalycin Structural Motif

The common structural motif seen in the X-ray structures for RBP, BBP, INSEC, MUP, and BLG has been described as a β-barrel or β-clam (Godovac-Zimmermann, 1988; Peitsch and Boguski, 1991). The β-barrel core is a single globular domain composed of two β-sheets stacked orthogonally to one another; RBP is shown as an example in figure 2–9. The principal feature of this core is the 8 anti-parallel β-strands labelled A to H which fold with a simple $(+1)_7$ topology (Cowan *et al.*, 1990). This topology is shown diagramatically in figure 2–10. The hydrogen bonding between strands results in two orthogonally stacked sheets, some strands being shared between sheets. In the case of RBP the front sheet is composed of strands ABCDEF, while the back sheet is composed of strands EGFHA and a ninth C-terminal strand I (Cowan *et al.*, 1990). The sharing of strands is made possible by the presence of β-bulges in the shared strands. An α-helix is present between the 8th and 9th strands. This packs against the external side of the back sheet, interactions being mainly hydrophobic in character. In RBP the helix is seen to pack at an optimal angle of approximately -15° to the β-sheet (Cowan *et al.*, 1990). All structures solved so far show at least one disulphide bond (table 2–1). Superposition of the structures shows that RBP, BLG and MUP share a common disulphide bond between residues 70-174, 66-160 and 64-157 respectively. Neither BBP or INSEC have a disulphide bond structurally equivalent to this.

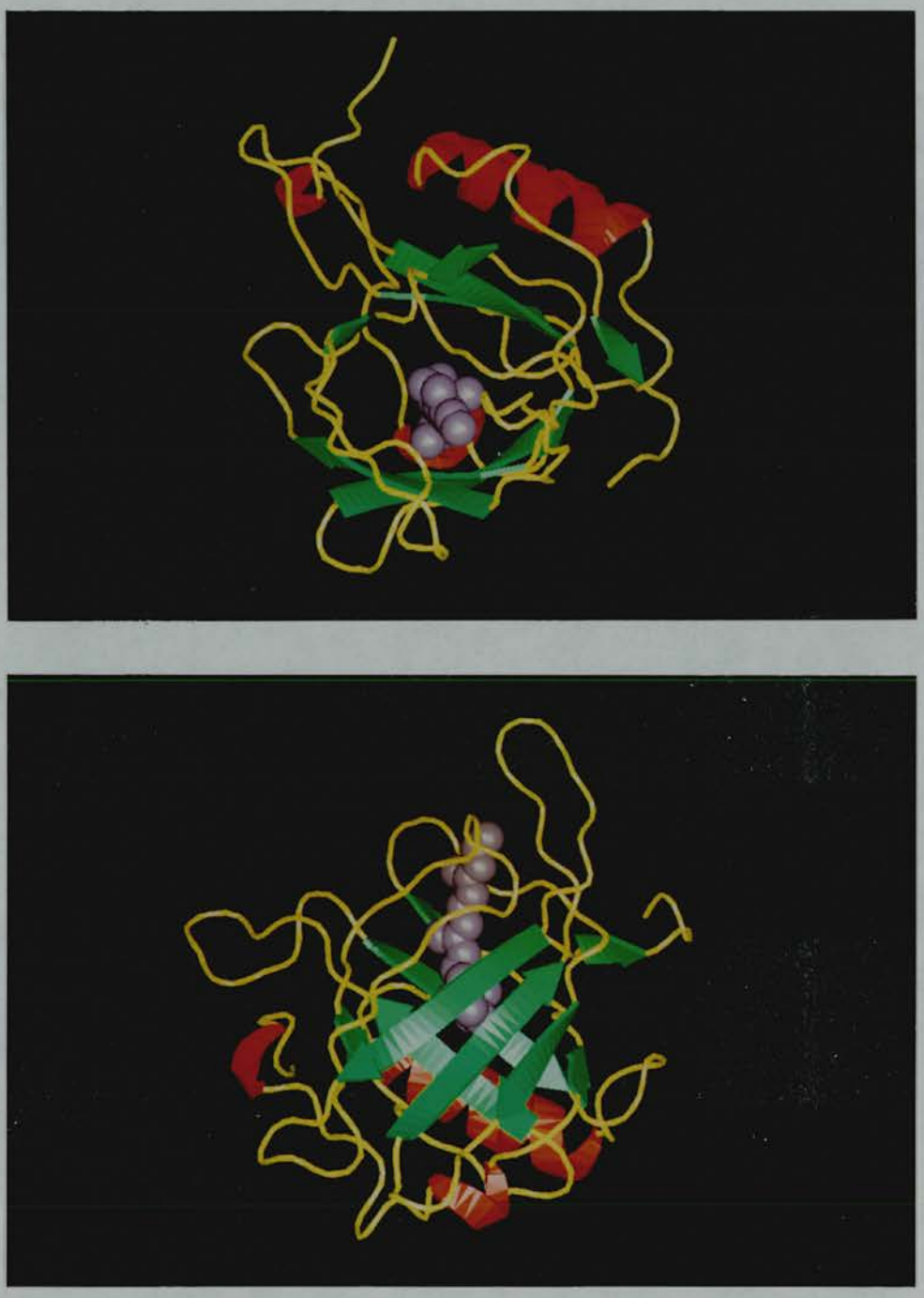


Figure 2–9: Cartoon representation of human plasma RBP, top: view along the binding calyx, bottom: view perpendicular to the stacked β-sheets. Beta strands are represented by green arrows, alpha helices by red helices. The purple space-filled molecule is bound retinol. Both pictures generated using sketch_auto in the program O.

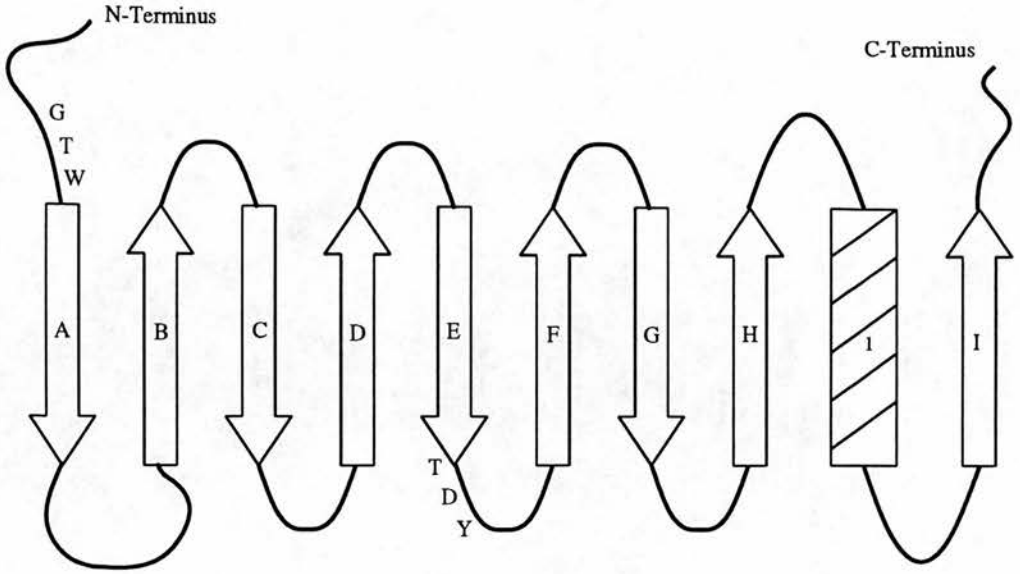


Figure 2–10: Schematic representation of the lipocalycin topology

2.5.2 Sequence Similarity Between Lipocalycins

Despite the high structural homology, between those members whose structures have been determined, sequence homology is generally seen to be low between the lipocalycins. Indeed, Dayhoff *et al* present the alignment of bovine BLG and A1GP as being at the limits of detection of sequence homology (Dayhoff *et al.*, 1983). Sequence identity scores range from 60% and above (66% for homologous proteins such as a2u and MUP) down to below 10% (7% identity for RBP and A1GP). The assignment of proteins to this family is mainly on the basis of two conserved sequence motifs. These motifs are (using the a2u numbering scheme):

$$G_{17}\text{-x-}W_{19} \text{ and } T_{95}\text{-D-}Y_{97}$$

The tryptophan at residue 19 is the only absolutely conserved residue in the set of sequences identified so far. Both these regions are conserved structurally in the structures of RBP, BBP, BLG, INSEC and MUP. Structural superposition of these motifs and the surrounding residues (-2,+2) was carried out using the program O (T. A. Jones, University of Uppsala). Residues were aligned explicitly using the lsq_explicit command; all alignments were made to RBP. The sequence alignments of these small regions show absolute conservation of G_{17}, W_{19} and

```
rbp      FSGTWYA          rbp      VDTDYDT
mup      INGEWHT          mup      PKTDYDN
bbp      YHGKWWE          bbp      LATDYKN
insec    FAGAWHE          insec    LSTDNKN
         . * *                     . **
```

Figure 2–11: Sequence alignment of the conserved lipocalycin G-x-W motif and T-D-Y motif.

	MUP	BBP	INSEC
RBP	0.419	0.366	0.306
MUP	-	0.648	0.570
BBP	-	-	0.362

Table 2–2: RMS deviations in position of superimposed alpha carbon atoms for G-x-W motif.

T_{95}-D_{96} (figure 2–11), although only W_{19} is conserved throughout the whole family. The three dimensional superposition of these motifs is extremely good as can be seen from the rms deviations between equivalent alpha carbon atoms (table 2–2 and table 2–3). The identity also extends to side chain conformation with non-conserved side chains also lying in equivalent positions (figure 2–12 and figure 2–13). The rigid body alignment of the structures shows similar tertiary structure (figure 2–14) despite low sequence similarity (figure 2–15). All sequences apart from BBP, INSEC and apolipoprotein-D (apoD) have cysteine residues at positions equivalent to 64 and 157 in a2u. A structural analysis of conserved residues is presented in chapter 4.

	MUP	BBP	INSEC
RBP	0.197	0.256	0.154
MUP	-	0.380	0.126
BBP	-	-	0.315

Table 2–3: RMS deviations in position of superimposed alpha carbon atoms for T-D-Y motif.

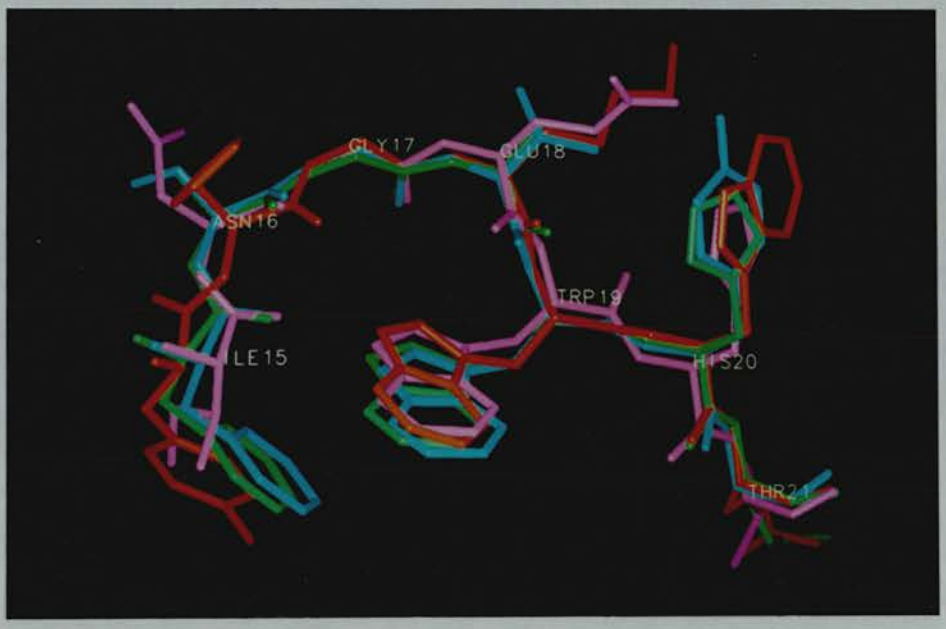


Figure 2–12: Superimposed residues in G-x-W motif for RBP (cyan), MUP (magenta), BBP (red) and INSEC (green).

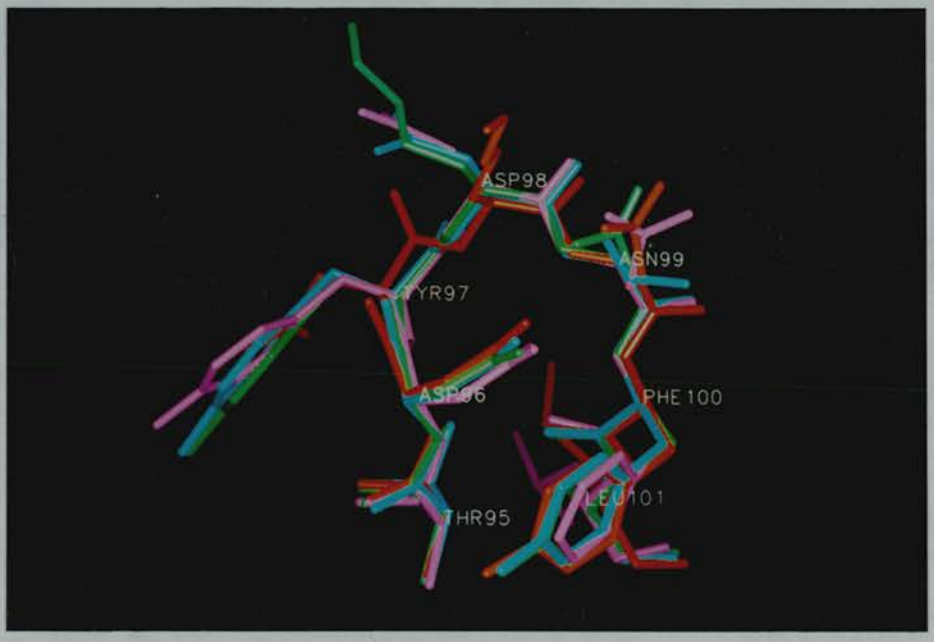


Figure 2–13: Superimposed residues in T-D-Y motif for RBP (cyan), MUP (magenta), BBP (red) and INSEC (green).

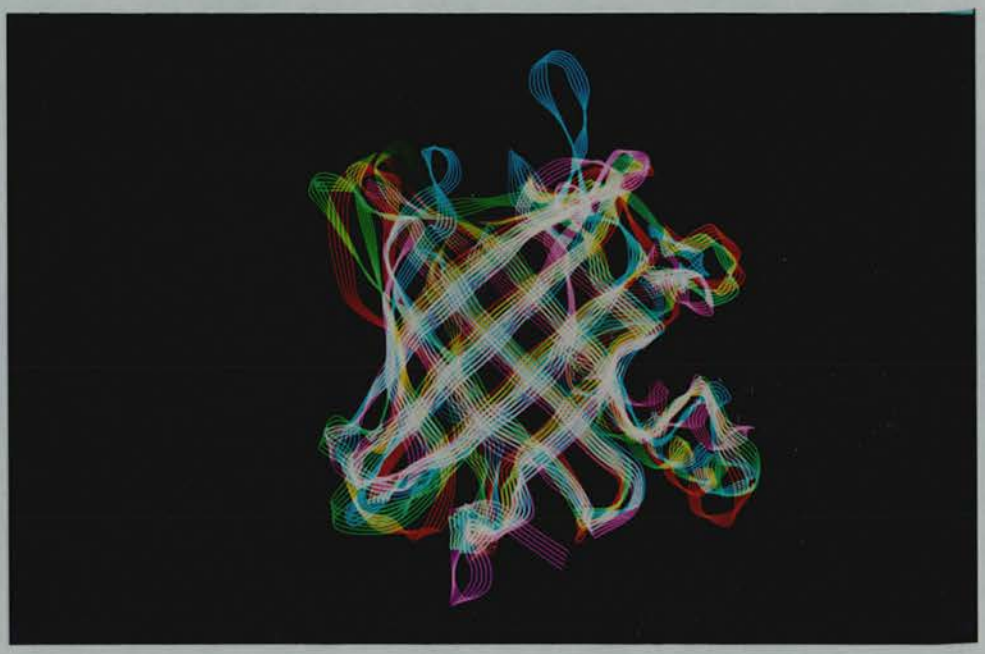


Figure 2–14: Rigid body alignment of RBP (cyan), MUP (magenta), BBP (red) and INSEC (green), ribbon representation.

2.6 Functional Role of the Lipocalycins

The lipocalycin family can be classified into several groups on the basis of biological activity and sequence homology. These groups will be discussed with relevance to specific proteins. The properties of the proteins are summarised in table 2–4.

2.6.1 Urinary Proteins

A high protein content in urine (proteinuria) is usually associated with kidney damage. However, male rats and both male and female mice excrete a large quantity of protein in the urine, the presence of which is not normally associated with pathological changes in the kidney. The physiological function of the a2u, MUP and proteins with related amino acid sequences will be discussed here.

```
>rbp       ------------------------------ERDCRVSSFRVKENFDKARFSGTWYAMAK
>purp      SSGSPAPLPNRMKYAQYVFLASIFSAVEYSLAQTCAVDSFSVKDNFDPKRYAGKWYALAK
>blg       -----------------------------------LIVTQTMKGLDIQKVAGTWYSLAM
>pp14      -----------------------------------MDIPQTKQDLELPKLAGTWHSMAM
>a2u       -----------------------------------EEASSTRGNLDVAKLNGDWFSIVV
>mup       -----------------------------------EEASSTGRNFNVEKINGEWHTIIL
>24p3      -------------------QSQAQDSTQNLIPAPSLLTVPLQPDFRSDQFRGRWYVVGL
>a2urel    -------------------QRQAQDSTQNLIPAPPLISVPLQPGFWTERFQGRWFVVGL
>epid      ---------------------------------------AVVKDFDISKFLGFWYEIAF
>pgds      ----------------------------GFPQTPAQGHDTVQPNFQQDKFLGRWYSAGL
>ch21      -----------------------------------AATVPDRSE-----VAGKWYIVAL
>a1gp      -----------MALSWVLTVLSLLPLLEAQIPLCANLVPVPITNAT-LDQITGKWFYIAS
>a1mg      ------MRSLGALLLLLSACLAVSAGP-----VPTPPDNIQVQENFNISRIYGKWYNLAI
>c8g       ----MLPPGTATLLTLLLAAGSLGQKPQRPRRPASPISTIQPKANFDAQQFAGTWLLVAV
>pbsn      ------------MRVILLLLTL-------DVLGVS--SMMTDKNL-KKKIEGNWRTVYL
>aph       ---------------------------------------------QDF-AE-LQGKWYTIVI
>veg       --------------------MKALLLTFGLSLLAALQAQAFPTTEENQDVSGTWYLKAA
>bg        --------------------MIRIIAIVVLFFLQCQADLPPVMKGLEENKVTGVWYGIAA
>pbp       ----------------------------------AQ--EEEAEQNL-SE-LSGPWRTVYI
>obp       -------------------MVKFLLIVLALGVSCAH--HENLDISP-SE-VNGDWRTLYI
>insec     ----------------------------GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAK
>bbp       ----------------------------NVYHDGACPEVKPVDNFDWSNYHGKWWEVAK
>apod      ---------MVMLLLLLSALAGLFGAAEGQAFHLGKCPNPPVQENFDVNKYLGRWYEIEK
                                                                 * *

>rbp       KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWD--VCADMVGTFTDTEDPAKFK
>purp      KDPEGLFLQDNISAEYTVEEDGTMTASSKGRVKLFGFWV--ICADMAAQYTVP-DPTTPA
>blg       AASDISLLDAQSA-PLRVYVEELKPTPEGDLEILLQKWENGECAQKKIIAEKTKIPAVF-
>pp14      ATNNISLMATLKA-PLRVHITSLLPTPEDNLEIVLHRWENNSCVEKKVLGEKTGNPKKF-
>a2u       ASNKREKIEENGSMRVFMQHIDVL-ENSLGFK--FRIKENGECRELYLVAYKTPEDGEY-
>mup       ASDKREKIEDNGNFRLFLEQIHVL-ENSLVLK--FHTVRDEECSELSMVADKTEKAGEY-
>24p3      AGNAVQ-KKTEGSFTMYSTIYELQENNSYNVTSILVRDQDQGCRYWIRTFVPSSRAGQF-
>a2urel    AANAVQ-KERQSRFTMYSTIYELQEDNSYNVTSILVRGQ--GCRYWIRTFVPSSRPGQF-
>epid      ASKMGTPGLAHKEEKMGAMVV--ELKENL-LALTTTYYSEDHCVLEKVTATEGDGPAKF-
>pgds      ASNSSWFREKKELLFMCQTVVAPSTEGGLNLTSTFLRKNQ--CETKVMVLQPAGVPGQY-
>ch21      ASNTDFFLREKGKMKMVMARISFLGEDELEVS--YAAPSPKGCRKWETTFKKTSDDGEL-
>a1gp      AFRNEEYNKSVQEIQATFFYFTP-NKTEDTIFLREYQTRQDQCIYNTTYLNVQRENGTIS
>a1mg      GSTCPWLKKIMDRMTVSTLVLGEGATEAE-ISMTSTRWRKGVCEETSGAYEKTDTDGKF-
>c8g       GSACRFLQEQGHRAEATTLHVAPQGTA---MAVSTFRKLDGICWQVRQLYGDTGVLGRF-
>pbsn      AASSVEKINEGSPLRTYFRRIEC-GKRCNRINLYFYIKKGAKCQ-QFKIVG-RRSQDVYY
>aph       AADNLEKIEEGGPLRFYFRHIDC-YKNCSEMEITFYVITNNQCS-KTTVIGYLKGNGTYQ
>veg       AWDKEIPDKKFGSVSVTPMKI-KTLEGGN-LQVKFTVLIAGRCKEMSTVLEKTDEPAKY-
>bg        ASNCKQFLQMKSDNMPAPVNIYSLNNGHMKSSTSFQT--EKGCQQMDVEMTTVEK-GHY-
>pbp       GSTNPEKIQENGPFRTYFRELVF-DDEKGTVDFYFSVKRDGKWK-NVHVKATKQDDGTYV
>obp       VADNVEKVAEGGSLRAYFQHMEC-GDECQELKIIFNVKLDSECQ-THTVVGQKHEDGRYT
>insec     LPLENENQGKCTIAEYKY--DGKKASVYNSFVSN----GVKEYMEGDLEIAPDAKYTKQG
>bbp       YPNSVEKYGKCGWAEYTP--EGKSVKVSNYHVIH----GKEYFIEGTAYPVGDSKI---G
>apod      IPTTFEN-GRCIQANYSLMENGKIKVLNQELRAD----GTVNQIEGEATPVNLTEPAK--
                                                          c
```

Figure 2–15: Multiple sequence alignment of members of the lipocalycin superfamily. Sequences from the GenBank (release 70.0) and EMBL (release 29.0). A lower case letter c marks the conserved cysteine residues. Continued overleaf

```
>rbp       ---MKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGL
>purp      KMYMTYQGLASYLSSGGDNYWVIDTDYDNYAITYACRSLKEDGSCDDGYSLIFSRNPRGL
>blg       ----KIDALN------ENKVLVLDTDYKKYLLFCMENSAEPEQSLACQ---CL-VRTPEV
>pp14      ----KINYTV------ANEATLLDTDYDNFLFLCLQDTTTPIQSMMCQ---YL-ARVLVE
>a2u       ----FVE----YDGGNTFTIL--KTDYDRYVMFHLINFKN-GETFQLM---VLYGRTKDL
>mup       ----SVT----YDGFNTFTIP--KTDYDNFLMAHLINEKD-GETFQLM---GLYGREPDL
>24p3      ----TLGNMHRYPQVQSYNVQVATTDYNQFAMVFFRKTSENKQYFKI----TLYGRTKEL
>a2urel    ----TLGNIHSYPQIQSYDVQVADTDYDQFAMVFFQKTSENKQYFKV----TLYGRTKGL
>epid      ------QVTRLSGKKE---VVVEATDYLTYAIIDITSLVAGAVHR-TM---KLYSRSLDD
>pgds      ----TYNSPHWQLPLP----LSVETDYDEYAFLFSKRTKGPGQDFRMA---TLYSRAQLL
>ch21      ----YYS----EEAEKTVEVL--DTDYKSYAVIFATRVKD-GRTLHMM---RLYSRSREV
>a1gp      ---RYVGGQE------HFAHLLILRDTKTYMLAFDVNDE-KNWGL------SVYADKPET
>a1mg      ----LYHKSKWNITME---SYVVHTNYDEYAIFLTKKFSRHHGPTITA---KLYGRAPQL
>c8g       ----LLQARGARGAVH---VVVAETDYQSFAVLYL-----ERAGQLSV---KLYARSLPV
>pbsn      ---AKYEGST------AFMLKTVNEKILLFDYFNRNRRNDVTRVA------GVLAKGRQL
>aph       ---TQFEGNN------IFQPLYITSDKIFFTNKKNMDRAGQETNM------IVVGKGNAL
>veg       ----TAYSGK---QVL---YIIPSSVEDHYIFYYEGKIHRHHFQ--IA---KLVGRDPEI
>bg        ---------KWKMQQGDSETIIVATDYDAFLMEF-TKIQMGAEVCVTV---KLFGRKDTL
>pbp       ---ADYEGQN------VFKIVSLSRTHLV-AHNINVDKHGQTTEL------TELFVKLNV
>obp       ---TDYSGRN------YFHVLKKTDDIIFF-HNVNVDESGRRQCD------LVAGKREDL
>insec     ---KYVMTFKFGQRVVNLVPWVLATDYKNYAINYNCDY-HPDKKAHSIHAWILSKSKV-L
>bbp       ---KIYHKLTYGGVTKENVFNVLSTDNKNYIIGYYCKY-DEDKKGHQDFVWVLSRSKV-L
>apod      ------LEVKFSWFMPSAPYWILATDYENYALVYSCTC-IIQL-FHVDFAWILARNPN-L

>rbp       PPEAQKIVRQRQEELCLAR---QYRLIVHNGYCD--GRSERNLL-------
>purp      PPAIQRIVRQKQEEICMSG---QFQPVLQSGAC-----------------
>blg       DDEALEKFDKALKALPMHIR-LSFNPTQLEEQCHI---------------
>pp14      DDEIMQGFIRAFRPLPRHLW-YLLDLKQMEEPCRF---------------
>a2u       SSDIKEKFAKLCEAHGITRD-NIIDLTK-TDRC--LQARG----------
>mup       SSDIKERFAQLCEEHGILRE-NIIDLSN-ANRC--LQARE----------
>24p3      SPELKERFTRFAKSLGLKDD-NIIFSVP-TDQC--IDN------------
>a2urel    SDELKERFVSFAKSLGLKDN-NIVFSVP-TDQC--IDN------------
>epid      NGEALYNFRKITSDHGFSET-DLYILKH-DLTC--VKVLQSAAESRP----
>pgds      KEELKEKFITFSKDQGLTEE-DIVFLPQ-PDKC--IQE------------
>ch21      SPTAMAIFRKLARERNYTDE-MVAVLPSQEE-C--SVDEV----------
>a1gp      TKEQLGEFYEALDCLRIPKSDVVYTDWK-KDKCEPLEKQHEKERKQEEGES
>a1mg      RETLLQDFRVVAQGVGIPED-SIFTMAD-RGEC--VPGEQEPEPILIPRVR
>c8g       SDSVLSGFEQRVQEAHLTED-QIFYFPK-YGFC--EAADQFH--VLDEVRR
>pbsn      TKDEMTEYMNFVEEMGIEDENVQ--RVMDTDTCPNKIRIR----------
>aph       TPEENEILVQFAHEKKIPVENIL--NILATDTCPE---------------
>veg       NQEALEDFQSVVRAGGLNPD-NIFI-PKQSETCPLGSN------------
>bg        PEDKIKHFEDHIEKVGLKKE-QYIR-FHTKATCVPK--------------
>pbp       EDEDLEKFWKLTEDKGIDKKNVV-------NFLENEDHPHPE---------
>obp       NKAQKQELRKL-----------------------------------------
>insec     EGNTKEVVDNVLKTFSHLIDASKFISNDFSEA--ACQYSTTYSLTGPDRH-
>bbp       TGEAKTAVENYLIG-SPVVDSQKLVYSDFSEA--ACKVN------------
>apod      PPETVDSLKNILT--SNNIDVKKMTVTD--QV--NCPKLS----------
```

Protein	Size	Source	Ligands
α-2u-globulin	162 aa	plasma, kidney, urine	pheromones?
Major urinary protein	162 aa	plasma, kidney, urine	pheromones?
Aphrodisin	151 aa	hamster vaginal discharge	?
a2u-related protein	198 aa	?	?
Plasma retinol binding protein	182 aa	plasma	retinol
Purpurin	196 aa	retina	retinol
Rat epididymal secretory protein	165 aa	epididymis	?
Bilin binding protein	189 aa	insect haemolymph	biliverdin IXγ
β-Lactoglobulin	162 aa	milk whey	retinol
Pregnancy Protein 14	162 aa	human endometrium	?
α_1-microglobulin	183 aa	plasma, urine	IgA, yellow-brown chromophore
α_1-acid glycoprotein	187 aa	plasma, urine	?
Complement protein C8γ	182 aa	plasma	retinol retinoic acid
Pyrazine binding protein	159 aa	bovine nasal mucus	pyrazine based odorants
Rat odorant binding protein	150 aa	nasal mucus	odorants?
Bowman's gland protein	160 aa	nasal mucus	odorants?
Von Ebner's gland protein	170 aa	saliva around taste buds	odorants?
Prostaglandin D synthase	168 aa	brain tissue	prostaglandin-H_2 prostaglandin-D_2
Apolipoprotein D	169 aa	plasma high density lipoprotein	cholesterol
Chondrocyte 21 protein	158 aa	skeletal tissue	?
Probasin	160 aa	Prostate epithelial cells	?

Table 2–4: Summary of properties of the lipocalycin family members.

Alpha-2u-Globulin

A biochemical and toxicological background for a2u has been given earlier in this chapter. Of interest here is the physiological function of a2u. The binding of small hydrophobic molecules by a2u has been studied by several workers. The binding of a variety of small aliphatic molecules, many substituted with oxygen or chlorine, has been investigated in the course of work on the nephropathy produced by a2u (table 2–5; Borghoff *et al.*, 1991; Lehman-McKeeman and Caudill, 1992). Sequence similarity to other members of the lipocalycin family involved in odorant binding prompted binding studies with a number of odorant molecules (table 2–5; Cavaggioni *et al.*, 1990). Other work has suggested that a2u binds fatty acids *in vivo* (Kimura *et al.*, 1989). Despite the extensive study of a2u in relation to hydrocarbon-induced nephropathy and ligand binding *in vitro* its physiological role still remains unclear. It is suggested that the protein functions as a pheromone carrier (Cavaggioni *et al.*, 1987). The hypothesis is that the protein is excreted into the urine carrying a bound pheromone molecule. The pheromone is released from the protein as the deposited urine dries denaturing the protein. Association of the volatile pheromone with a protein molecule would allow its slower release into the environment than excretion of the pheromone directly into the urine alone. This hypothesis is supported by the affinity of a2u for some small volatile odorant molecules *in vitro.* In addition, a proteinaceous fraction from male rat urine has been identified as having pheromonal properties (Vandoren *et al.*, 1983). It is probable that this function is only one of the physiological roles of a2u. The high rate of catabolism of a2u in the proximal tubule of the kidney is similar to that of RBP. The expression of a2u in different tissues and also its presence in low levels in the female rat suggests some role within the body, such as fatty acid transport (Kimura *et al.*, 1991).

Mouse Major Urinary Protein

The ligand binding properties of MUP have been less well studied than a2u. The nephrotoxic agent *d*-limonene-1,2-epoxide is seen not to bind to MUP *in vitro*

Compound	K_m (M)[a]
Methyldihydro-jasmonate	6.0×10^{-6}
Thymol	1.6×10^{-6}
2-Nonenal	3.0×10^{-6}
Cyneole	0.8×10^{-6}
(-)Methylfenchol	0.3
	K_a (M)[b]
[^{3}H]2,4,4-trimethylpentan-2-ol	5.6×10^{6}
	K_i (M)[b,c]
d-Limonene oxide	5.1×10^{-7}
2,4,4-Trimethyl-1-pentanol	6.8×10^{-7}
2,4,4-trimethylpentan-2-ol	7.6×10^{-7}
2,2,4-Trimethyl-1-pentanol	1.7×10^{-6}
α-Tetralone	5.4×10^{-6}
Isophorone	7.7×10^{-6}
Retinol	8.8×10^{-6}
α-Tetralol	7.7×10^{-5}
d-Limonene	1.0×10^{-4}
2-Hexanone	1.1×10^{-4}
2,5-Dichlorophenol	3.9×10^{-4}
Phenethyl alcohol	4.0×10^{-4}
1,4-Dichlorobenzene	5.2×10^{-4}

Table 2–5: Ligand binding data for a2u *in vitro*. [a]Data from Cavaggioni *et al.*, 1990. [b]Data from Borghoff *et al.*, 1991. [c]Apparent inhibition constant with respect to [^{3}H]2,4,4-trimethylpentan-2-ol binding to a2u.

Compound	K_m (M)[a]
Methyldihydro-jasmonate	8.0×10^{-6}
2-Nonenal	6.0×10^{-6}
(+)Methylfenchol	6.0×10^{-6}
β-Ionone	1.6×10^{-6}
Thymol	1.3×10^{-6}

Table 2–6: Ligand binding data for MUP *in vitro*. [a]Data from Cavaggioni *et al.*, 1990.

(Lehman-McKeeman and Caudill, 1992). That hydrocarbon-induced nephropathy is not seen in either male or female mice suggests that MUP is not capable of binding the same small hydrocarbons as a2u. However, the differences in renal biology between the rat and mouse may be sufficient to account for this observation. It is clear that MUP can bind some of the same odorant molecules as a2u (table 2–6; Cavaggioni *et al.*, 1990). MUP has also been purified from mouse urine with physiological ligands still bound (Bacchini *et al.*, 1992). Extraction of ligands from the protein identified 2-(sec-butyl)thiazoline (70%), 2,3-dehydro-exo-brevicomin (15%), and 4-(ethyl)phenol (15%). The first two compounds are known to have pheromonal activity when present in male rat urine. It should be noted that only 40% of the MUP contained a bound ligand; other physiological ligands are therefore possible. However, MUP is strongly implicated as a pheromone carrier molecule. It is suggested that MUP is excreted with a protease which is secreted from the bladder lining (Flannery *et al.*, 1990). This protease could proteolytically cleave MUP in the deposited urine, thus releasing the bound ligand (Beynon personal communication). This would provide a slow, controlled release of the volatile pheromone into the atmosphere. It is possible that this pheromone carrier activity of MUP is its major physiological role; no renal reabsorption is seen in either male or female mice. However, the synthesis of different MUP variants in different tissues (Shahan *et al.*, 1987a) may imply an internal physiological function such as ligand transport.

Aphrodisin

This is a soluble protein, molecular mass 17kD, which is the major protein component of vaginal discharge in hamsters (Henzel *et al.*, 1988). It is seen to produce a copulatory response in male hamsters. It is suggested that the protein itself acts as a pheromone (Henzel *et al.*, 1988). The protein sequence has many features in common with the lipocalycins suggesting it is a further member of the family. It is unclear whether ligand binding is one of the physiological roles of aphrodisin (APH); it seems unneccessary for its pheromonal activity.

Alpha-2u-Globulin Related Protein

A protein sequence, deduced from the nulceotide sequence, shows some similarity to the protein sequence of a2u and has been called a2u related protein (A2UREL) (Chan *et al.*, 1988). As no biochemical data are available for this protein its function is unknown. The sequence is almost identical to a protein 24P3 which is induced by SV40 infection in mice (Hraba-Renevey *et al.*, 1989). It is possible that these are one in the same protein.

2.6.2 Retinoid Binding Proteins

Retinol and other retinoids such as retinoic acid are extremely important biological molecules. Retinol is needed for vision, as the light sensitive component in rhodopsin. Retinoic acid has shown to be a key chemical in differentiation; human nuclear retinoic acid receptors have been identified (Giguere *et al.*, 1987). Many of the lipocalycins have been shown to bind retinol or retinoids *in vitro*, however, only plasma retinol binding protein has been shown to bind retinol *in vivo*.

Plasma Retinol Binding Protein

Retinol binding protein (RBP) has been identified as the major retinol transporter in plasma (Kanai *et al.*, 1968). The role of RBP is many fold. It

carries out the transfer of insoluble retinol between tissues, mainly from the liver to peripheral tissues. The strong association between the retinol and the protein serves to protect the retinol from oxidation and ensure its distribution is only to specific targets. It is also thought that RBP has a role in the transport of retinol from the maternal circulation to the developing foetus. Human plasma RBP is monomeric with a molecular mass of approximately 21 kD. It is seen to have high affinity and specificity for all-trans-retinol, its physiological ligand. The crystal structure of human plasma retinol binding protein shows the retinol bound within the cavity formed by the β-barrel. To fulfill its function as a retinol transporter RBP is associated with another plasma protein - transthyretin. Transthyretin is a tetramer of identical subunits giving a total molecular mass of 55 kD. The stoichiometry of the association between RBP and transthyretin remains unclear - two or four binding sites for RBP per dimer are suggested (Fex *et al.*, 1979). A putative binding site for transthyretin on RBP has been identified on the basis of invariant residues in different species and the structure of human plasma RBP (Cowan *et al.*, 1990). The complex is too large to be filtered into the urine. However, upon loss of the bound retinol it is thought that the apo-RBP dissociates from the transthyretin molecule, allowing the apo-RBP to be filtered into the urine. This urinary apo-RBP is then reabsorbed and catabolised in a manner similar to a2u.

Purpurin

The amino acid sequence of retina neural cell protein, purpurin (PURP), indicates that it is a member of the lipocalycin family (Berman *et al.*, 1987). High sequence similarity is seen to plasma RBP. The major physiological function of PURP is thought to be as a component in the extracellular adheron complex around retina cells (Schubert and LaCorbiere, 1985). It mediates the adhesion of retinal nerve cells to the cell-substratum by specific interaction with a cell surface receptor. However, it has also been shown to bind retinol (Schubert *et al.*, 1986). It is unclear what importance the binding of retinol has with respect to its cell adhesion function.

Rat Epididymal Secretory Protein

This protein is secreted into the lumen of the male rat epididymis under the control of androgens (Brooks *et al.*, 1986). Rat epididymal secretory protein (EPID) has been shown to bind retinoic acid with high affinity and specificity (Newcomer and Ong, 1990). The binding of retinoic acid is believed to be related to a role in sperm maturation (Newcomer and Ong, 1990).

2.6.3 Insect Pigment Proteins (Bilin Binding Proteins)

At least two proteins in the lipocalycin family act as colouring agents in insects. Their ability to colour comes from the interaction between the protein and a bound ligand. One of the major haem breakdown products in insects, biliverdin IXγ, is often seen to be associated with proteins. This protein/bilin complex is deposited in the epidermis or interlamellar space of the wing providing colouration. The physiological role of these bilin-binding proteins is assumed to be in pigmentation. Insecticyanin (INSEC) from tobacco hornworm (*Manduca sexta*) and bilin-binding protein (BBP) from the cabbage white butterfly (*Pieris brassicae*) have been studied in detail. The sequences and crystal structures of both proteins have been determined (Riley *et al.*, 1984; Zuber *et al.*, 1987; Holden et al., 1987; Huber *et al.*, 1987a,b). These proteins are 173 and 192 amino acids long respectively. The crystal structures suggest that both proteins exist as a tetramer *in vivo* and produce a blue pigmentation when complex with biliverdin IXγ. As reported earlier, comparison of the sequence of INSEC to other lipocalycins indicated that this was also a lipocalycin - a fact born out by the crystal structure. The crystal structure of BBP also showed the same tertiary structure, which is not unexpected given the moderately high level of sequence identity to INSEC (43%).

2.6.4 The Lactoglobulins

Beta-Lactoglobulin

Beta-lactoglobulin (BLG) is the major protein of whey from the milk of cows and many other mammals. The protein can be easily purified from milk in relatively large quantities. The BLGs have been extensively studied, producing a large amount of information (reviewed in Hambling *et al.*, 1991). However, the function of the protein still remains elusive. It is suggested that it may be a retinol transport protein, carrying retinol from the mother to the infant through the milk although this hypothesis has yet to be proved. It is possible that the protein functions as a general small lipophillic molecule transporter between mother and child. The sequence of BLG from many different species has been determined indicating a protein of 162 amino acids. The gene for bovine BLG has been expressed in *E.coli* and the gene for ovine BLG in yeast (Cho *et al.*, 1992; Paterson, 1991). The crystal structure of 3 different crystal forms of BLG have been solved; X, Y, and Z (Yewdall, 1988; Papiz *et al.*, 1986; Monaco *et al.*, 1987). The structure of the lattice Y form was solved by multiple isomorphous replacement (MIR), while the structure of both X and Z was solved by a combination of low resolution MIR and molecular replacement using the lattice Y structure. All three structures show the same tertiary fold with significant differences being in loop regions. No electron density is seen for a ligand in the lattice X and Y structures. In lattice Z ligand electron density is reported in an external region of the protein; at the interface between the helix and β-sheet (Monaco *et al.*, 1987). However, mutagenesis work suggests that retinol binds within the β-barrel calyx (Cho *et al.*, 1992). Other work using a combination of different ligands suggests that there are at least two different ligand binding sites in BLG (Dufor and Haertle, 1990; Dufour *et al.*, 1991).

Pregnancy Protein 14

This protein is secreted by the human endometrium between the late luteal phase of the menstrual cycle and the first trimester of pregnancy (Bell, 1986).

The production of the pregnancy protein 14 (PP14) is progesterone regulated. The monomer size is 162 amino acids, which is also glycosylated. The sequence of human PP14 has been determined (Julkenan *et al.*, 1988), and is seen to show high similarity to BLG. On this basis the protein has been classified as a lipocalycin. The biochemical role, or ligand binding capabilities of the protein have yet to be determined.

2.6.5 Immune Response Proteins

Response to invasion by foreign bodies in vertebrates is seen to be both at the molecular level (e.g. soluble antibodies) and at the cellular level (e.g. T-cells). Both α_1-microglobulin and α_1-acid glycoprotein have been identified as members of the lipocalycin family and also soluble macromolecules involved in the immune response. Protein C8γ is a component of the membrane disrupting complex of human complement; part of the cellular immune response system, and is also suggested to be a lipocalycin.

Alpha-1-Microglobulin

This is a a plasma protein of approximately 30 kD which can be detected in the urine of human patients with kidney disfunction (Ekstrom *et al.*, 1975). Alpha-1-microglobulin (A1MG) has a sequence length of 188 but is also glycosylated at three residues; two N-linked and one O-linked. The homologous protein has also been observed in other species (Akerstrom, 1985) many of which have been partially or completely sequenced (Akerstrom and Logdberg, 1990). The principal sites of sysnthesis are the liver and kidney. The function of the protein remains unclear, but it is suggested that a major role is in the regulation of the immune system (Akerstrom and Logdberg, 1990). Sequence comparison has indicated that the protein is a member of the lipocalycin family. Alpha-1-microglobulin can induce cell division of lymphocytes, but can also inhibit stimulation of lymphocytes by other protein antigens. Neutrophil granulocyte migration can be also be inhibited *in vitro* by A1MG. Cell surface

receptors for A1MG have been identified and may represent a mechanism for direct effects on cells. Alpha-1-microglobulin has a covalently linked chromophore, which imparts a yellow-brown colour to the free protein in solution (Akerstrom and Logdberg, 1990). This colavent attachment is to the free cysteine at residue 34 (Escribano *et al.*, 1991). Extraction of the purified protein with hexane suggests that another major ligand, amongst others, is retinol (Escribano *et al.*, 1988). The covalent ligand imparts a charge heterogeniety and colour to the protein. When the protein becomes complexed with plasma immunoglobulin A; about 50% of human plasma A1MG is, this charge heterogeneity and colour is lost. It is suggested that the covalent chromophore is involved in a specific covalent interaction with the plasma IgA acting as a cross-linking agent. The physiological importance of this interaction remains unclear.

Alpha-1-Acid Glycoprotein

Alpha-1-acid glycoprotein (A1GP) is an acute phase reactant plasma protein whose concentration is seen to increase during infection or inflamation. It is thought the protein acts as a general immunosuppressor but this is still not clear. The protein is 183 amino acids in length but is heavily glycosylated with 5 N-linked sugar chains. Binding activity includes many cationic ligands and some therapeutic drugs (Kremer *et al.*, 1988). The sequence of the protein suggests it is a lipocalycin.

Complement Protein C8γ

The complement system is responsible for the lysis of foreign cells, preparation of foreign cells for phagocytosis, and generation of peptides which regulate the immune response. The complex range of roles is reflected in a large number of protein components; at least 25. The lysis of foreign cells occurs as the end product of a complex cascade mechanism. The end product of this cascade is a large protein complex which is able to form pores in the foreign cell membrane - leading to the lysis of the cell (Muller-Eberhard, 1986). The complex cascade

mechanism can be activated by both the classic and the alternative complement pathway. The latter stages of the pathway have component C5β interacting with component C8. This complex then binds with many monomers of component C9 which polymerises into a polyC9 membrane pore. The C8 component is itself multimeric; consisting of two glycosylated proteins, C8α and C8β, and a non-glycosylated protein, C8γ. The function of C8γ remains unclear, it is not essential for the cell lysis activity of the complex. Instead the protein may play a part in the protection of hosts cells from lysis by interaction with host cell surface receptors (Sodetz, 1988; Hansch, 1988). The sequence homology to A1MG possibly indicates a similarity in their mode of action through interaction with cell surface receptors. This sequence similarity suggests that C8γ is a lipocalycin; a fact supported by its ability to bind retinol and retinoic acid. As with A1MG the physiological importance of this binding is unclear.

2.6.6 Olfactory Proteins

The molecular basis of olfaction (smell) would seem to have to be necessarily complex. Relatively recently specific proteins have been discovered associated with olfactory tissue which can may be involved in odorant recognition (Margolis 1987; Snyder *et al.*, 1988). The sequences of some of these proteins suggest a relationship to other lipocalycins. This is perhaps not unexpected, if a2u and MUP bind and transport pheromone molecules it is not unreasonable to assume that the receptors or odorant sequestering proteins in the olfactory system could bind the pheromones in a similar way; a protein homologous to a2u or MUP could bind the same pheromone.

Bovine Pyrazine Binding Protein

A protein has been isolated from the nasal mucosa of cows which has a high affinity for a known odorant molecule; 3-isobutyl 3-methoxy pyrazine (Pevsner *et al.*, 1985; Bignetti *et al.*, 1985). One odorant molecule is bound per two 19kD protein molecules, suggesting the active unit is a dimer (Pevsner *et al.*, 1985).

Immunological studies show this pyrazine binding protein (PBP) to be localised mainly to the epithelium of the nasal mucosa, although it is also present in the olfactory mucosa and neurons and nasal secretions (Pevsner *et al.*, 1985). The sequence of the protein shows lipocalycin like features (Cavaggioni *et al.*, 1987; Tirindelli *et al.*, 1989), although no disulphide bridges are present.

Rat Odorant Binding Protein

A homologous protein to bovine pyrazine binding protein has been cloned and sequenced from rat nasal mucosa (Pevsner *et al.*, 1988). The protein, odorant binding protein (OBP), is located in the largest of the discrete nasal glands of the rat, Sterno's gland. The amino acid sequence is similar to that of PBP and is therefore thought to also be a lipocalycin. It is interesting to note that rat OBP has several cysteines which could participate in disulphide bonds, whereas PBP has no cysteine residues at all.

Bowman's Gland Protein

A protein has been identified, by cloning and sequencing, from the Bowman's gland in frog nasal tissue (Lee *et al.*, 1987). The Bowman's gland protein (BG) appears to be localised to the olfactory tissue and its sequence shows similarity to the two previous odorant binding proteins. Therefore, it is suggested that this protein is a frog odorant binding protein and also a member of the lipocalycin family.

Von Ebner's Gland Protein

The von Ebner's glands secrete saliva in the region of the taste buds. A small protein of 18kD has been cloned and sequenced from the acinar cells of this gland in male rats (Schmale *et al.*, 1990). The sequence shows similarity to other lipocalycins, specifically the odorant binding proteins. A analogy can be drawn with the olfactory mucosa; von Ebner's gland protein (VEG) fulfills a similar role

in ligand binding and transport to receptors. However, no biochemical data has been produced to support this idea.

2.6.7 Other Lipocalycins

There are lipocalycins which do not seem to fall into any specific subgroup althought the binding of hydrophobic ligands is still implicated for some of them. The first protein described below is the only lipocalycin for which an enzymatic activity has been described to date (Peitsch and Boguski, 1991).

Prostaglandin D Synthase

This protein is responsible for the conversion of prostaglandin-H_2 to prostaglandin-D_2 in the brains of both rats and humans (Nagata *et al.*, 1991). Prostaglandin D is synthesised in other tissues by a glutathione dependent enzyme which is distinct from brain prostaglandin D synthase (PGDS) (Urade *et al.*, 1985). Prostaglandin D is important physiologically in sleep induction, body temperature regulation, anticonvulsion, suppression of luteinising hormone release, pain perception and odour responses (Hayaishi, 1988). The rat enzyme is 168 amino acids long with two N-linked glycosylation sites. The sequence of both human and rat enzymes have been determined and show highest homology to A1MG and C8γ (Nagata *et al.*, 1991). PGDS differs with respect to the latter two proteins in being permenantly membrane associated, although it is noted that both A1MG and C8γ do interact with other proteins at membrane surfaces as part of their physiological function. The sequence information indicates that PGDS is a lipocalycin. It is suggested that PGDS contains a free cysteine at residue 65, equivalent to Phe45 in RBP. This cysteine could make contact with the bound prostaglandin by analogy with RBP and retinol. Sulfhydryl compounds are known to be vital for the enzyme's activity, implicating this residue in the catalytic mechanism (Nagata *et al.*, 1991).

Apolipoprotein D

Apolipoprotein D (apoD) is a minor component of high density lipoprotein particles (HDLs) in mammalian plasma (McConathy and Alaupovic, 1973). Its interaction with the other lipoproteins is known to be important but its biological function still remains unclear. The mature protein is 169 amino acids long, with glycosylation at one or two asparagine residues (Drayna et al., 1986). The sequence shows the greatest homology to both insecticyanin and bilin binding protein, a fact which prompted the modelling of the tertiary structure of apoD (Peitsch and Boguski, 1990). The modelling results suggest that apoD is suited to the binding of bilin like molecules. This contradicts the view that apoD is responsible for the binding and transport of cholesterol esters (Fielding and Fielding, 1980). It is also suggested that apoD may bind progesterone and pregnenolone (Simard *et al.*, 1991).

Chondrocyte 21 Protein

This protein of 21 kD can be used as a marker for the stage of differentiation of mesenchymal cells during bone formation. The protein is secreted by cultured chondrocytes (bone forming cells) as they differentiate, and is present in most skeletal tissue with a time dependent distribution (Cancedda *et al.*, 1988, 1990). The amino acid sequence of condrocyte protein 21 (Ch21) has been determined (Cancedda *et al.*, 1990) and shows homology to a2u and MUP. A protein from cultured chicken heart mesenchymal cells, quiescence specific protein (QSP), is seen to be identical in sequence to Ch21 (Berard *et al.*, 1987, 1989). It is suggested that the proteins plays a role in stabilising mature cell populations (Cancedda *et al.*, 1990). The method of action of the protein remains unknown.

Probasin

Another protein whose abundance seems to be linked to the state of cell differentiation is probasin (PBSN). It has a molecular mass of 20 kD and a very basic pI of 11.5 (Matuo *et al.*, 1984). The protein is located in the nuclei of the

prostate epithelial cells and is also secreted (Spence *et al.*, 1989). The function of the protein is unknown, but the amino acid sequence shows some homology to the urinary and odorant binding proteins, suggesting that it is a lipocalycin (Spence *et al.*, 1989).

2.7 Proteins with Lipocalycin-like Folds

The lipocalycin superfamily is composed of proteins which share common sequence motifs and, it is inferred, a common tertiary structure. However, they are not unique in possessing an antiparallel β-barrel structure. Several proteins have been identified whose tertiary structures show similarity to the lipocalycin fold but do not possess those lipocalycin sequence motifs.

2.7.1 Fatty Acid Binding Proteins

A family of intracellular low molecular mass proteins has been identified which can bind fatty acids and other hydrophobic ligands (Ockner, 1990). Two major sub-groups have been identified within this family; fatty acid binding proteins (FABP) and cellular retinoid binding proteins. The FABPs can be subdivided into three distinct types; liver, intestinal and heart (cardiac). It is thought that the FABPs have a role in transport and storage of fatty acids within the cell. The cellular retinoid binding proteins include a retinol binding protein (cRBP) and a retinoic acid binding protein (cRABP). These retinoid binding proteins are thought to be involved in cytoplasmic transport of their ligands between organelles, and maybe most importantly to the nucleus where retinoids are thought to have regulatory effects. Other members of the FABP family fall outside these two main groups. However, those identified to date have apparent roles in fatty acid binding and transport. The tertiary structures of four members of the FABP family have been determined; P2-myelin protein (Jones *et al.*, 1988), rat intestinal FABP (Sacchettini *et al.*, 1988), chicken liver FABP (Scapin *et al.*, 1990), and bovine heart FABP (Muller-Fahrnow *et al.*, 1991). The

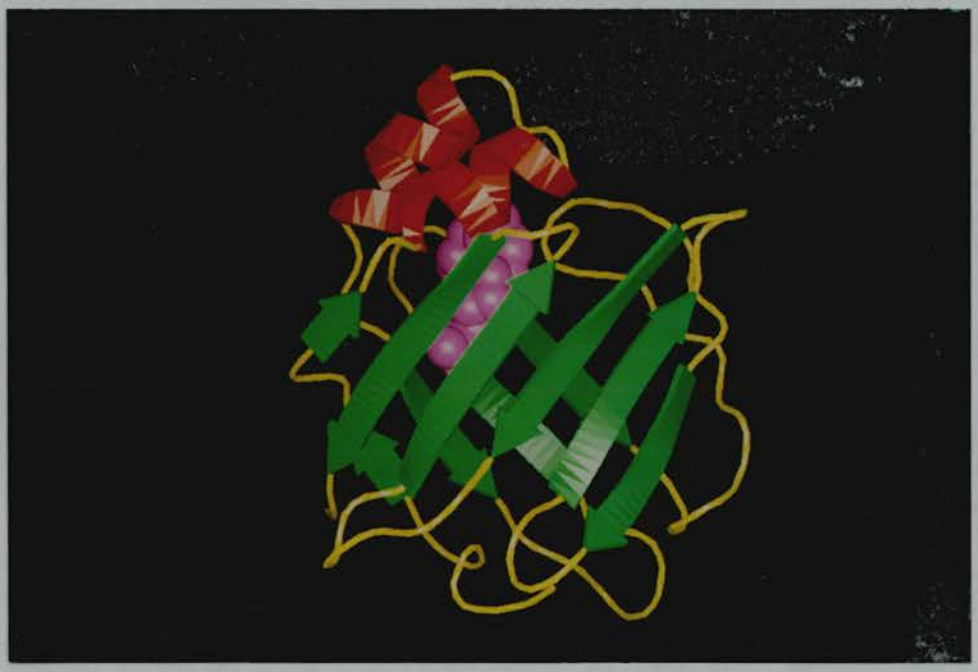


Figure 2–16: Cartoon representation of P2-myelin protein. Beta strands are represented by green arrows, alpha helices by red helices. The purple space-filled molecule is the bound fatty acid.

coordinates for P2-myelin protein and rat intestinal FABP are available; both show a lipocalycin like fold with anti-parallel β-strands forming a β-barrel. The major difference to the lipocalycin structure is in the number of strands; there are 10 instead of the 8 observed in the lipocalycins. The conserved sequence motif G-x-W is present in the FABP family but not the T-D-Y motif. The interaction between the conserved tryptophan and a basic residue (Arg122 in a2u) is also preserved in the FABP family (Jones *et al.*, 1988). The lipocalycin α-helix is not conserved, but two shorter helices are present between the first and second β-strand. These helices interact with the bound fatty acid molecule in P2-myelin protein (figure 2–16). The best conserved structural regions between P2-myelin protein and human plasma RBP are the G-x-W motif and strands G and H. This low level of structural identity and the low levels of sequence similarity between the FABPs and the lipocalycins justifies their partitioning into two distinct families.

2.7.2 Cyclophilin

Cyclophilin is the major intracellular receptor for the immunosuppressive drug cyclosporin A (Handschumacher *et al.*, 1984). The latter acts as an inhibitor of T-cell activation and can prevent graft rejection in organ and bone marrow transplants. Cyclophilin may be responsible for mediating this immunosuppressive response. The protein is also a peptidyl-prolyl isomerase; catalysing the interconversion between the *cis* and *trans* isomers of proline in both protein and peptide substrates (Fischer *et al.*, 1989). Human cyclophilin is a single polypeptide chain of 165 amino acids. Its active substrate cyclosporin A is a cyclic undecapeptide of fungal origin. The structure of human cyclophilin has been determined by X-ray crystallography and NMR spectroscopy (Kallen *et al.*, 1991). The overall fold is seen to be an eight-stranded antiparallel β-barrel. There are also two α-helices, neither of which is equivalent to the lipocalycin α-helix. The topology of this barrel differs from the very simple one of the lipocalycins; instead of having a $(+1)_7$ topology cyclophilin has a more complex +1, -3, -1, -2, +1, -2, -3 topology. In addition the β-barrel core of cyclophilin is packed with hydrophobic residues hindering internal ligand binding; instead the putative ligand binding site is at the surface of the protein on the outside of one of the β-sheets (Kallen *et al.*, 1991). These structural differences and the low sequence identity with any lipocalycin suggest that cyclophilin is not a lipocalycin. A protein with similar immunosuppressant activity to cyclophilin, FK-binding protein, has been sequenced and its tertiary structure determined by both NMR and X-ray crystallography. Both methods show a five stranded antiparallel β-sheet wrapped around a short helix (Michnick et al., 1991; Van Duyne *et al.*, 1991). Therefore, the tertiary fold of cyclophilin is not neccessarily a common feature of prolyl isomerases or immunosuppressive proteins.

2.7.3 Streptavidin

Streptavidin (STVN) is a protein from the bacteria *Streptomyces avidinii* with a high affinity for the vitamin biotin. The structure of the complex of STVN and

biotin has been solved by X-ray crystallography (Weber *et al.*, 1989). The structure of a truncated STVN (core streptavidin) has also been solved independently (Hendrickson *et al.*, 1989). An eight-stranded β-barrel is reported in both cases, prompting the suggestion of a relationship between STVN and the lipocalycin and FABP families (Hendrickson *et al.*, 1989). However, no significant sequence identity between STVN and any lipocalycin or FABP is reported. Unfortunately coordinates for STVN are not available to study the possible structural similarities with the lipocalycins. It is reasonable to assume that STVN is not a member of either the lipocalycin or FABP families.

2.7.4 Catalase

Catalase is a liver enzyme which converts hydrogen peroxide into water and molecular oxygen. It is a single polypeptide chain, multi-domain protein of 500 residues. The structure of bovine liver catalase has been determined by X-ray crystallography (Murthy *et al.*, 1981). An eight-stranded antiparallel β-barrel can be seen as part of one domain. The topology of this barrel is complex but does have the same overall shape as the lipocalycin fold. The non-significant sequence identity between catalase and the lipocalycins, and their vastly different biological functions suggest that catalase is not a lipocalycin.

2.7.5 Photoactive Yellow Protein

Photoactive yellow protein (PHY) is isolated from the bacterium *Ectothiorhodospira halophila.* This purple autotrophic bacteria lives in salt water and uses reduced sulphur compounds as electron donors in photosynthesis (McRee *et al.*, 1986). The protein PHY is a single polypeptide chain of 126 residues which is yellow-coloured. It displays a reversible photocycle similar to that of sensory rhodopsin. The structure of PHY has been determined by X-ray crystallography (McRee *et al.*, 1989). Only the α-carbon coordinates are available for this structure at present. Visual inspection of the structure shows an antiparallel ten-stranded sandwich, five β-strands in each half of the

sandwich. The strand topology is $(+1)_9$ but the overall shape is not a barrel; it is flattened to form a box-like structure. This difference in structure and the lack of sequence similarity to the lipocalycins suggests that PHY is not a lipocalycin.

Chapter 3

X-ray Crystallography

3.1 Background

The determination of molecular structure at the atomic level is possible by either X-ray crystallography or high-field nuclear magnetic resonance (NMR). Both techniques are well documented and reviewed elsewhere (Blundell and Johnson, 1976; Stout and Jensen, 1968; Wüthrich, 1986), therefore what follows is a brief background to crystallography. This is followed by the solution of two small molecule structures as an example. Finally there is an account of the work leading to X-ray diffraction analysis of a2u and MUP.

3.1.1 X-rays and the Unit Cell

X-rays are scattered by the electrons which orbit the nucleus of an atom. The interference between the X-rays scattered from the atoms in a structure produces significant changes in the intensity of diffracted waves observed in different directions. This variation arises because the path differences taken by the scattered X-rays are of the same magnitude as the separation of atoms in the structure. X-rays are used because they have a wavelength similar to interatomic distances in structures (typically 1.5 Å). It is not possible to study the diffraction of X-rays from one molecule in isolation. Instead, a crystal of many such molecules is studied. A crystal is a regular, repeating array of atoms or molecules in three dimensions. Crystals are usually described in terms of a lattice, which is a geometric construction defined by three axes and the three angles between

them. The lengths of these axes, labelled a, b, and c, denote where a repeat of the molecule occurs. The angles between b and c, a and c, and a and b are labelled α, β, and γ respectively. This parallelepiped described by lengths a, b and c and angles α, β and γ is called the unit cell. It can be thought of as containing all the atoms that contribute to the diffraction of incident X-rays. The unit cell can only be of 14 different basic types (called Bravais lattices). Although the basic unit of the crystal is this unit cell there can be symmetry relationships between the molecules within the unit cell. The various symmetry operations possible produce 32 distinct point groups (symmetry operations which leave at least one point, in the object to which they are applied, unmoved). The convolution of these 14 lattices and 32 point groups produces 230 unique space groups for three dimensional crystals. The space group can be identified from the diffraction pattern by the systematic absence of reflections.

3.1.2 How X-rays Interact with Crystals

If we consider three planes of atoms each separated by distance d and incident and reflected X-rays at angle θ to these (figure 3–1). For the reflected waves to be in phase the path difference must be an integer number of wavelengths. If this is the case then the reflected waves will interfere constructively, this is Bragg's Law, which can be expressed:

$$n\lambda = 2d\sin\theta \tag{3.1}$$

Provided there are a large number of planes contributing to this diffraction the position in space at which a given reflection is observed is well defined. The position at which these reflections occur is defined by the crystal lattice, while the scattered intensity is dependent on the electron density and therefore the arrangement of atoms in the cell. Refering back to equation 3.1 the order of diffraction is defined by n, this can be considered in two ways. For a given separation d there are higher orders of diffraction which occur at larger scattering angles. Alternatively, scattering can be thought of as arising from

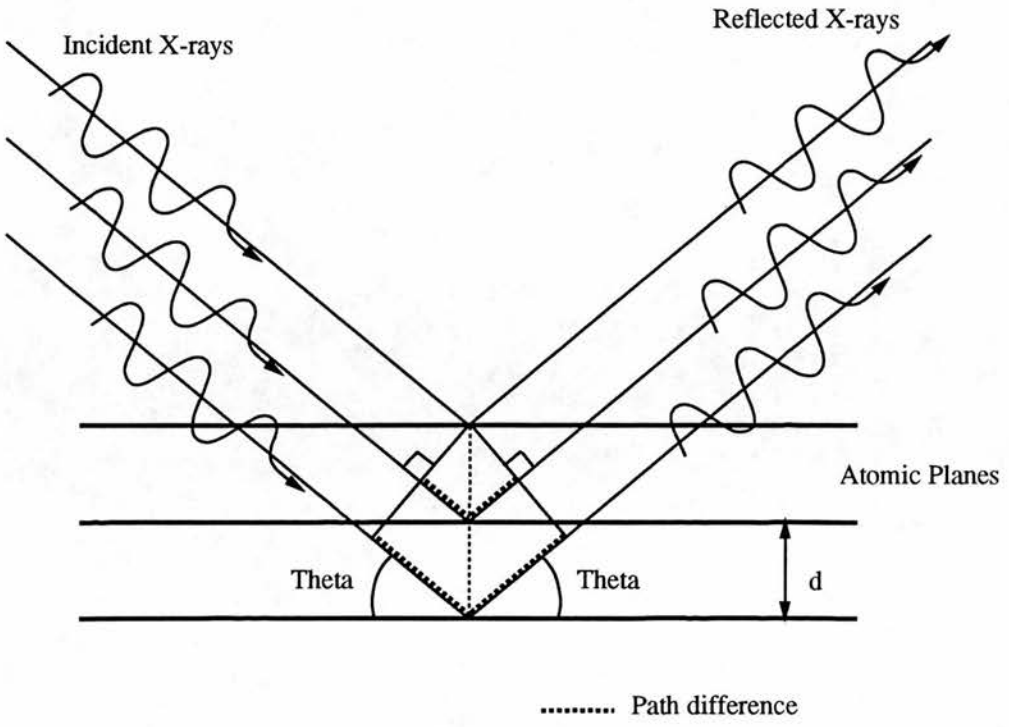


Figure 3–1: Geometric conditions for X-ray diffraction

planes which are closer together. Therefore a reflection at a given value of θ can be the nth order from planes at spacing d, or as the first order from planes of spacing d/n. The diffraction pattern from a protein crystal therefore has many reflections, some method for indexing these must be used. The Miller indices are most commonly used, which uses the order of diffraction with respect to each of the unit cell axes. The symbols h, k and l are used for a, b and c respectively. Bragg's Law gives information about the geometric conditions for diffraction but says nothing about the relationship between the diffracted rays and the way atoms are arranged in the unit cell, this is given by the Fourier transform.

$$F_{hkl} = \sum_{j}^{atoms} f_j \exp\left[2\pi i \left(hx_j + ky_j + lz_j\right)\right] \tag{3.2}$$

Where F_{hkl} is the complex structure factor, for reflection hkl, which consists of the amplitude and the phase of the diffracted ray. The atomic scattering factor for each atom j is f_j, and the fractional coordinates of each atom j are x_j, y_j,

and z_j. The inverse Fourier transform of the structure factors gives the electron density ρ at every point xyz within the unit cell.

$$\rho_{xyz} = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl} \exp\left[-2\pi i \left(hx + ky + lz\right)\right] \tag{3.3}$$

Where V is the unit cell volume. It can be seen therefore that each reflection has some contribution from every atom in the crystal, and the electron density is calculated with a contribution from every reflection. The aim of the crystallographer is to determine the structure factors for a crystalline array of the molecule of interest and hence determine the electron density for the unit cell. This would be relatively easy if the analysis of diffracted X-rays from a crystal gave the complete complex structure factor. Unfortunately diffracted X-rays only give information about the amplitude component of the structure factors since what is measured is the scattered intensity (the amplitude squared), phase information is lost. Therefore, much effort goes into determining the phase component of the structure factor, in general the larger the molecule the more difficult this task becomes. For small molecule (10-200 atoms) diffraction data, it is usually possible to solve the phases of reflections directly using the Patterson function or direct methods. These cannot be used for large molecules such as proteins, therefore isomorphous heavy atom derivatives need to be generated.

3.1.3 X-ray Sources

A collimated source of X-rays of a single wavelength is required for most diffraction studies. These are generated in the laboratory by accelerating a beam of electrons into an anode, the metal of which dictates the wavelength of the resulting X-rays. The excitation of the metal electrons and their return to ground state emits X-rays, in the case of a copper anode this wavelength is 1.5414 Å: the CuKα absorption edge. If the copper target is stationary the maximum electron beam power is limited due to heating and damage of the copper target. This is the case with the sealed beam X-ray source, which has a maximum power load of 2-3 kW on a target area of approximately 0.4 x 12 mm. In the rotating anode

source the copper target is rotated thus spreading the heating effects over a much larger surface. This allows a higher power load, resulting in a more intense X-ray beam. In order to obtain X-ray beams much more intense than the rotating anode a synchrotron source must be used. Electrons are accelerated around a circular ring at close to the speed of light. The work done in accelerating the electrons around bends results in the emission of radiation, at the appropriate bending angle and electron velocity X-rays are produced. For small molecule crystallography the sealed tube is preferred. For macromolecular crystallography a rotating anode is usually neccessary, and for very large unit cells or small crystals synchrotron radiation is often needed. There are advantages and disadvantages for all these X-ray sources. Sealed tube sources are cheap and simple to maintain but cannot produce an intense beam. Rotating anode sources are more expensive to purchase and maintain but have a significantly more intense X-ray output. Synchrotron sources are only available as a national resource, but the X-rays produced can be tuned to different wavelengths and are up to at least 1000 times more intense than a rotating anode source. The high intensity of the synchrotron radiation reduces the time for data collection but results in instability in the X-ray beam, the intensity of which decreases as the electrons lose velocity. These effects have to be considered in data processing.

3.1.4 Obtaining Crystals

The growth of protein crystals has been reviewed by several authors (McPherson, 1982; Ducruix and Giegé 1992). Different techniques exist, all with the same aim: to bring a protein solution slowly to a solubility minimum at which crystals may grow. Most often some kind of precipitant is used, such as ammonium sulphate. In the hanging drop method a drop of protein plus precipitant (usually 10 μl in total) is suspended above a well of precipitant. The system is sealed and slowly the drop and well equilibrate, with conditions in the drop moving closer to those of the precipitant in the well. If the precipitant concentration in the drop is initially too low to precipitate the protein but moves closer to this point slowly the protein may crystallise. The microdialysis

technique achieves the same result by slowly dialysing the protein solution against a more concentrated solution. The growth of crystals is dependant on many factors such as pH, temperature, ionic strength, and both protein and precipitant concentration. In most cases the conditions for crystallisation have to be determined by trial and error. It is also important to remember that the quality and source of the protein is important. Highly purified protein which is homogeneous both in molecular mass and charge is preferred.

3.1.5 Crystal Analysis

In order to examine the diffraction pattern of a protein crystal it must be moved about in a beam of X-rays. As it moves reflections will appear as Bragg's Law is satisfied for those reflections at the incident X-ray beam angle θ. Therefore, the intensity of a reflection can only be recorded when the geometrical arrangement of the X-ray beam , crystal orientation, and detector satisfies Bragg's Law. In order to collect diffraction data from a protein crystal an X-ray source, X-ray detector, and some method for moving the crystal in the X-ray beam is required. The X-ray source remains fixed, some methods depend on the X-ray detector moving some require it to be static. Of primary importance is some method for keeping the protein crystal in a stable environment during data collection.

3.1.6 Crystal Mounting

Typically crystals are mounted in fine glass capillary tubes. Enough moisture must remain in the tube such that the crystal does not dry out but not so much that the crystal slips during data collection. This is achieved by placing some small volume of the mother liquor at either end of the capillary before sealing with wax. During data collection it is possible to keep the temperature around this capillary constant by blowing a stream of cooled nitrogen over it. More recently the mounting of crystals has been changed by the technique of flash cooling. This technique emerses the crystal in a viscous oil which is then rapidly frozen in liquid nitrogen. The low temperature is maintained during data

collection by blowing a stream of liquid nitrogen over the crystal. The benefits of this technique are a reduction in radiation damage, and reduced loss of water of crystallization by the crystal during data collection. All protein crystals suffer radiation damage in an X-ray beam, due to the production of free-radicals. The rate at which this damage becomes apparent depends on the protein and often the particular crystal. If a protein crystal is damaged quickly by X-rays many different crystals will have to be used to collect a complete set of diffraction data.

3.1.7 Crystal Characterisation by Photographic Methods

It is desirable to determine the cell dimensions and symmetry of a crystal before data collection. The classical techniques commonly used to do this are Weissenberg and precession photography. Both methods are well described in other texts (Stout and Jensen, 1968). Briefly, Weissenberg photography produces a distorted image of the reflections from one plane of reciprocal space, for example the $h0l$ plane (figure 3–2). From the photographs obtained it is possible to measure the unit cell lengths and the angles between them. It is possible to take photographs of higher level planes, such as $h1l$, in order to determine the crystallographic symmetry. The precession technique record the same information as the Weissenberg technique but the diffraction pattern produced is not distorted (figure 3–3). The latter method is commonly used for analysing protein crystals prior to data collection. Both the Weissenberg and precession technique are used for characterisation of small molecule crystals. It is noted that the Weissenberg technique, using an electronic detector is used for protein crystal data collection at the Photon Factory in Japan.

3.1.8 Data Collection by Photographic Methods

The most commonly used method for protein crystal data collection is the rotation method (Arndt and Wonacott, 1977). The crystal is aligned with a real axis parallel to the incident X-ray beam. The crystal is then rotated back and forth (oscillated) around this axis. Many different reflections are brought into a

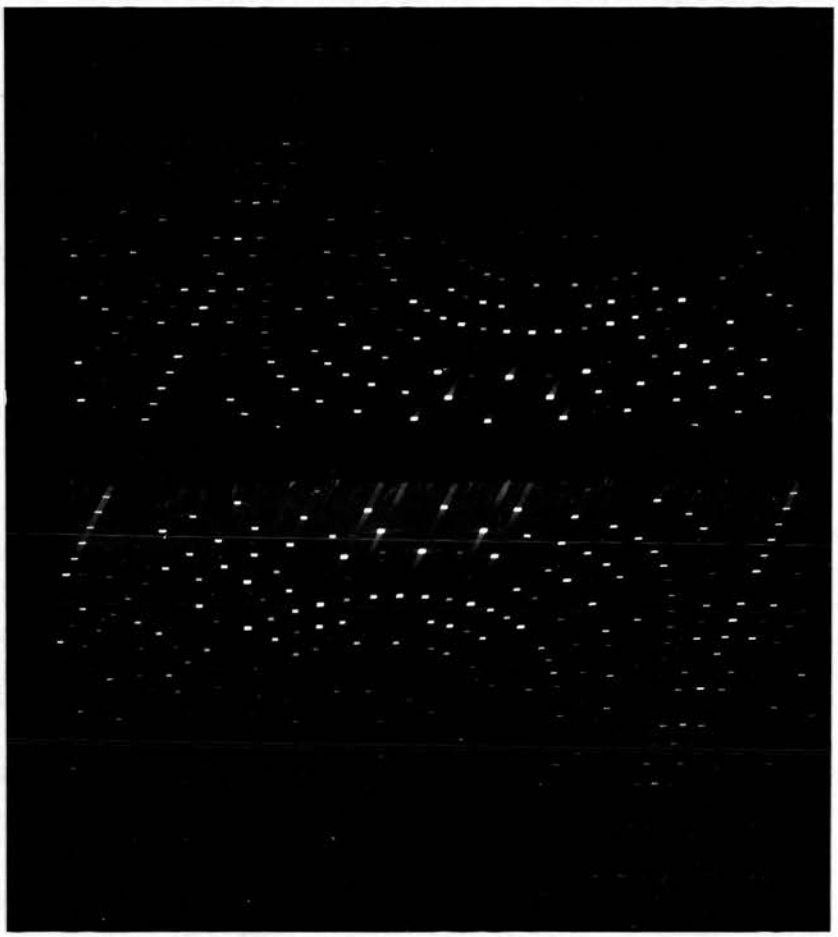


Figure 3–2: Weissenberg photograph of $h0l$ zone for NAN-190.

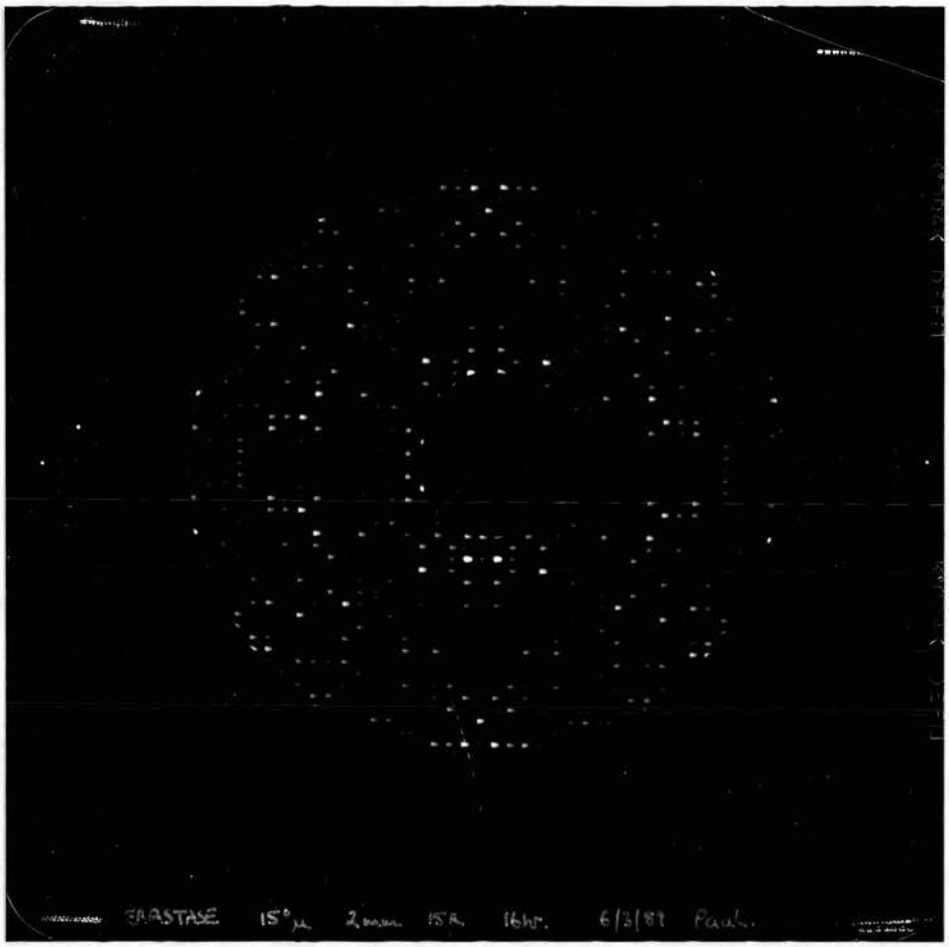


Figure 3–3: Precession photograph of $h0l$ zone of elastase.

diffracting position simultaneously. These reflections can be recorded on a film held in a flat or V-shaped cassette, films being changed between oscillations. Oscillation ranges are typically of the order of one or two degrees. By incrementing the start angle of oscillation all reflections to a certain resolution can be collected. The films are later digitised and processed to give the intensity of each reflection.

3.1.9 Data Collection using Electronic Techniques

The diffractometer consists of an X-ray (photon) proportional counter and a cradle for positioning of the crystal which together can measure single reflections very accurately. The crystal and detector can be moved so as to measure the intensity of different reflections one at a time. Since the crystal cell dimensions and their orientation with respect to the diffractometer must be known before data can be collected, it is possible to measure specific reflections of known index. This has the advantage that data are collected for reflections whose index is already known, therefore very little post-data collection processing is required. This method of measuring X-ray data is accurate but inefficient as only one reflection is measured at a time. This is the preferred method for small molecules which are usually stable, but macromolecules usually have a limited exposure life making the diffractometer of limited use. In addition the unit cells for macromolecules are usually large, dramatically increasing the number of reflections to be measured. The diffractometer can be used for the collection of low resolution protein diffraction data, often heavy atom derivative data, where accuracy of intensity measurements is of importance. The preferred method for collection of protein X-ray diffraction data is the rotation method with a two dimensional electronic detector of some kind. With fluorescent screen devices, such as the Enraf-Nonius FAST television diffractometer, many still images can taken with small steps between them, these can be processed later to give integrated intensities. Alternatively use of an image plate allows full oscillation images to be taken, as with the photographic technique, and processed in the same way as digitised film images.

3.1.10 Practical Examples

It can be seen from the above that the methods used to collect diffraction data for small and large molecules differ widely, although the aim is the same for both. The solution of two small molecule structures is given as an illustration of the work involved. This is followed by preliminary diffraction studies of a2u and crystallisation studies of MUP.

3.2 Structure Solution of Two Small Molecules

A series of 5-hydroxytryptamine receptor antagonists was synthesised and crystallised by I. Dawson (Department of Pharmacology, Edinburgh of University). The structures of two of these compounds were determined by X-ray crystallography.

3.2.1 Pharmacology of 5HT Receptors

Serotonin (5-hydroxytryptamine, 5HT) is synthesized from the amino acid tryptophan. Serotonin acts as a chemical messenger acting at the postsynaptic side of neurons. In general it is synthesized in nerve terminals where it is held in storage granules or vesicles, and released into the synaptic cleft at nerve impulse. Serotonin is formed by some non-neuronal cells such as the entero-chromaffin cells in the gut, and is stored and released, but not synthesized, by cells such as blood platelets. There appear to be serotonin receptors at many locations in peripheral tissues that apparently do not have any serotonin producing neurons. The role of these receptors is generally poorly understood. The neuronal serotonin system is better understood: the mammalian brain possesses an expansive serotonin producing neuronal circuitry. Many different areas of the brain are innervated by serotonin neurons, suggesting involvement in numerous brain functions.

Serotonin receptors occur in brain and in various peripheral tissues, including the gut, the uterus, the heart, and blood vessels. Serotonin receptors may be presynaptic or postsynaptic. Presynaptic receptors may modulate the release of other neurotransmitters. The classification of serotonin receptors has been based on the effects of different agonists and antagonists. Later work with radiolabelled chemicals has allowed a more detailed classification. At least 5 different groups have been identified: 5HT-1A,1B,1C,1D and 5HT-2. Of interest here is the 5HT-1A group which is thought to mediate certain behavioural effects, the hypothermic effect, and the antihypertensive effects of some serotonin agonists (Peroutka *et al.*, 1986).

There is much interest in modulating the activity of the serotonergic system. Serotonergic drugs may be useful in treating some psychiatric disorders, such as mental depression, anxiety, and alcoholism. It is clear that certain chemicals with a high affinity for 5HT-1A receptors show an anxiolytic affect (Wander *et al.*, 1986; Csanalosi *et al.*, 1987). Chemicals with 5HT-1A affinity can also increase food uptake (Dourish *et al.*, 1986), lower blood pressure (Martin and Lis, 1985), and show some antimigraine properties (Hiner *et al.*, 1986).

Of pharmacological interest is the synthesis of specific antagonists, and also radiolabelled ligands with very high specificity for 5HT-1A receptors. Few antagonists specific to 5HT-1A receptors have been reported: propranolol and pindolol being two. The design and synthesis of series of high-affinity 5HT-1A ligands has been described (Glennon *et al.*, 1988a). One of these agents, 1-(2-methoxyphenyl)-4-[4-(2-phthalimido)butyl]piperazine.HBr (NAN-190) binds with high affinity to 5HT-1A hippocampal sites in the brain. It is seen to have 5HT-1A antagonist activity against the 5HT-1A agonist 8-hydroxy-2-(di-*n*-propylamine)tetralin (8-OH-DPAT) (Glennon *et al.*, 1988b). An iodo-derivative, 1-(2-methoxyphenyl)-4-[4-(4-I-benzamido)butyl] piperazine.HBr (IMD-1) was synthesised as part of a series of compounds based around NAN-190. The intended use of such a compound would be as a specific marker for 5HT-1A receptors. This could be achieved by synthesis with I^{125} rather than I^{127}. This radiolabelled compound could be used to identify the location of

5HT-1A receptors in live specimens. Both NAN-190 and IMD-1 were synthesised and crystallised by I.Dawson (Dept. of Pharmacology).

The three-dimensional structure of these molecules is required to understand which chemical features of these molecules determine their agonist or antagonist activity.

3.2.2 Data Collection

In both cases the crystals were stable enough to be mounted on a glass fibre using Araldite glue. The crystals were first characterised using 10° oscillation photographs, aligning the longer crystal axis with the rotation axis. The other two axes were measured and crystal symmetry deduced using Weissenberg photography. The crystals were then transferred to a Siemens-Stöe AED-2 4-circle diffractometer. A molybdenum sealed tube was used, producing X-rays of wavelength 0.7109 Å. Strong reflections were found and auto-indexing used to determine unit cell dimensions and an orientation matrix. This unit cell was checked against that determined previously. Strong reflections at higher resolution were collected to refine the unit cell and orientation matrix and a subset of these reflections was used as intensity standards, to monitor any crystal decay during data collection. Unique data were then collected, the orientation matrix having been refined, and standard reflections measured every two hours. At the end of data collection a ψ scan, around a reciprocal lattice vector, was collected and in order to apply an empirical crystal absorption correction (North *et al.*, 1968). The data were then corrected for crystal decay with respect ot time. The indexed reflection intensities and estimated standard deviations in measurement were written to disk.

3.2.3 Structure Solution

The solution of the structure requires that the phases of the reflections also be known as well as the amplitudes. For some small molecule structures the

intensity measurements are accurate enough to obtain phase information from the distribution of intensities. The solution of phases by this direct method is often possible for centro-symmetric structures, as phases are either 0° or 180°. Alternatively, the presence of repeated features in the unit cell can lead to repeated features in the reflection intensities. These features can be studied using the Patterson method. The Patterson function is defined as;

$$\rho_{uvw} = \frac{1}{V} \sum_h \sum_k \sum_l |F_{hkl}|^2 \cos\left[2\pi i \left(hu + kv + lw\right)\right] \tag{3.4}$$

If one atom (or a very few atoms) is significantly heavier than the majority of atoms in the unit cell, for example bromine in a molecule otherwise containing only carbon, nitrogen, oxygen and hydrogen atoms, the Patterson synthesis can reveal the position of this heavy atom. The Patterson function at a point uvw can be thought of as the convolution of the electron density at all points xyz in the unit cell with the electron density at points $x + u$, $y + v$, $z + w$. If any two atoms in the unit cell are separated by vector uvw a peak in the Patterson function will occur at uvw. If the two atoms are heavier than other atoms in the unit cell this peak will be large. The Patterson function can often be used to locate the position of a heavy atom in the small molecule unit cell. The position of this atom can then be used to calculate phases (with eqn. 3.2). These phases are used as a first approximation for calculation of the electron density in the unit cell using the observed structure factor amplitudes (with eqn. 3.3). Both structures were solved using the Patterson method in SHELX-76 or SHELX-86 (Sheldrick, 1986). This was possible due to the presence of a bromide ion, as a salt of crystallisation in both crystals and the substituted iodine atom in IMD-1. Difference Fourier syntheses were then used to find peaks of electron density, to which atoms were assigned. Atom types could be assigned and the connectivity deduced on the basis of distance between atoms, a highly automated process in SHELX-76. Once all non-hydrogen atoms were assigned, hydrogen atoms were included and constrained to be 1.08 Åfrom their respective non-hydrogen atoms. Atomic positions and anisotropic temperature factors were refined using a least squares difference Fourier technique, again within SHELX-76. Hydrogen atoms

	NAN-190	IMD-1
Crystal dimensions (mm)	0.2x0.2x0.6	0.4x0.05x0.6
Space group	C2/c	$P2_1/a$
a (Å)	21.9177 (0.0020)	7.4401 (0.0040)
b (Å)	15.1977 (0.0034)	15.6073 (0.0012)
c (Å)	14.0434 (0.0017)	21.9117 (0.0011)
β (°)	101.559 (0.010)	93.506 (0.0050)
F_{000}	2008	1184
Calculated density (gcm^{-3})	1.401	1.549
θ_{max}	23.5	22.5
Maximum hkl	24 17 15	8 16 23
μ (cm^{-1})	18.04	28.3
Reflections measured	3541	3476
Merging R-factor	0.0 (Unique data)	0.0 (Unique data)
Absorption correction	Psi scan	Psi scan

Table 3–1: Data collection parameters for NAN-190 and IMD-1

were refined riding upon the atoms to which they were attached. The crystal, data collection and structure solution parameters are summarised in table 3–1 and table 3–2 respectively.

3.2.4 Analysis of the Structures

The atomic coordinates for the non-hydrogen atoms in each structure are presented in table A–1 and table A–2. Although anisotropic temperature factors were refined for each non-hydrogen atom, only isotropic factors are given,

$$U_{eq} = \frac{1}{3}\sum_i \sum_j U_{ij} a_i^* a_j^* a_i a_j \tag{3.5}$$

these indicate the positional disorder for each atom. The bond lengths, bond angles and torsional angles, for non-hydrogen atoms in both structures were calculated using the program CALC (Gould and Taylor, 1983). The bond lengths in both structures are close to those expected (table A–3 and table A–4). The bond angles (table A–5 and table A–6) and torsion angles (table A–7 and table A–8) are also close to those usually observed. The piperazine ring nitrogen (N4) is protonated in both structures (figure 3–6). Both structures have a water molecule and a bromide ion as salts of crystallisation. The water molecule in

	NAN-190	IMD-1
Reflections observed with $(\frac{F}{\sigma(F)} > 2)$	2246	2781
Weighting $(1/[\sigma^2 F + gF^2])$ g=	unit	0.00069
R_w	0.0447	0.0532
S	3.441	1.485
Maximum shift/esd for final cycle	-0.088	-0.346
Minimum electron density (e/Å^3) in final difference Fourier	-0.4505	-0.639
Maximum electron density (e/Å^3) in final difference Fourier	0.3993	0.977
Number of parameters refined	305	308
Final R-factor	0.0441	0.0440

Table 3–2: Structure solution parameters for NAN-190 and IMD-1

NAN-190 lies on a special position (0.0, 0.2842, 0.25). There is an electrostatic interaction between the protonated N4 of the piperazine ring and the bromide ion (present as a salt of crystallisation) in both the NAN-190 and IMD-1 structures (figure 3–4). The crystal structure of NAN-190 is not stabilised by any specific interactions, rather by a complementarity of packing between molecules. The crystal structure of IMD-1 is stabilised by several interactions:

- Stacking of the benzamido rings. This is presumably favourable because of close proximity of delocalised electron systems.
- A hydrogen bond between N2 (in the benzamido group) and OW1 (the water oxygen) through hydrogen HN2.
- A hydrogen bond between OW1 and N1 (a nitrogen of the piperazine ring) through HW1 (a water hydrogen).

The piperazine rings of the two structures were superimposed manually using FRODO (Jones, 1978). The piperazine rings and the first butyl carbon are essentially identical (figure 3–5). However, the methoxyphenyl ring in IMD-1 is rotated approximately 5° anticlockwise with respect to NAN-190 when looking from the piperazine ring. In NAN-190 the methoxyphenyl ring and phthalimido

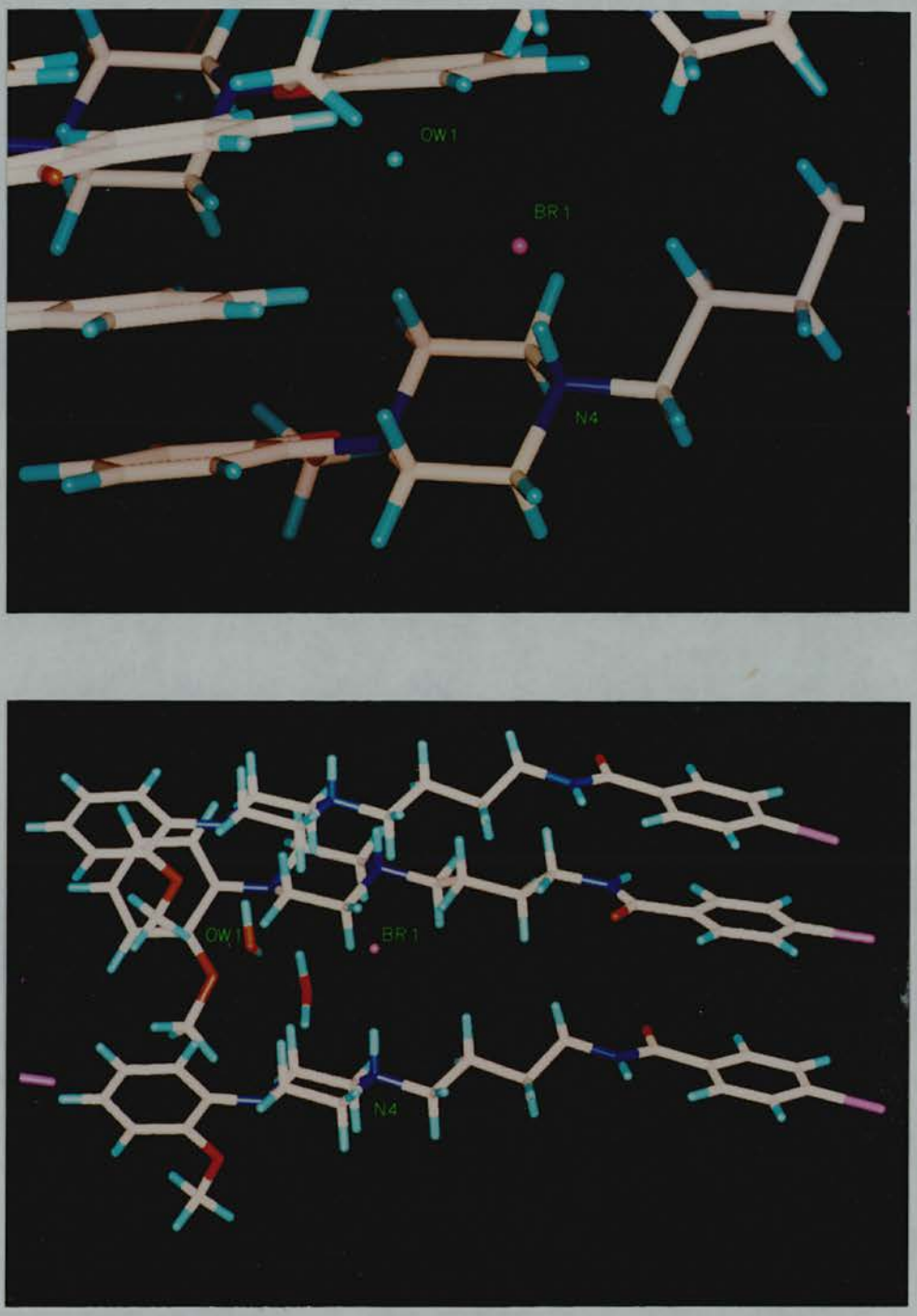


Figure 3–4: Interaction between N4 of piperazine ring and bromide ion in both NAN-190 (above) and IMD-1 (below).

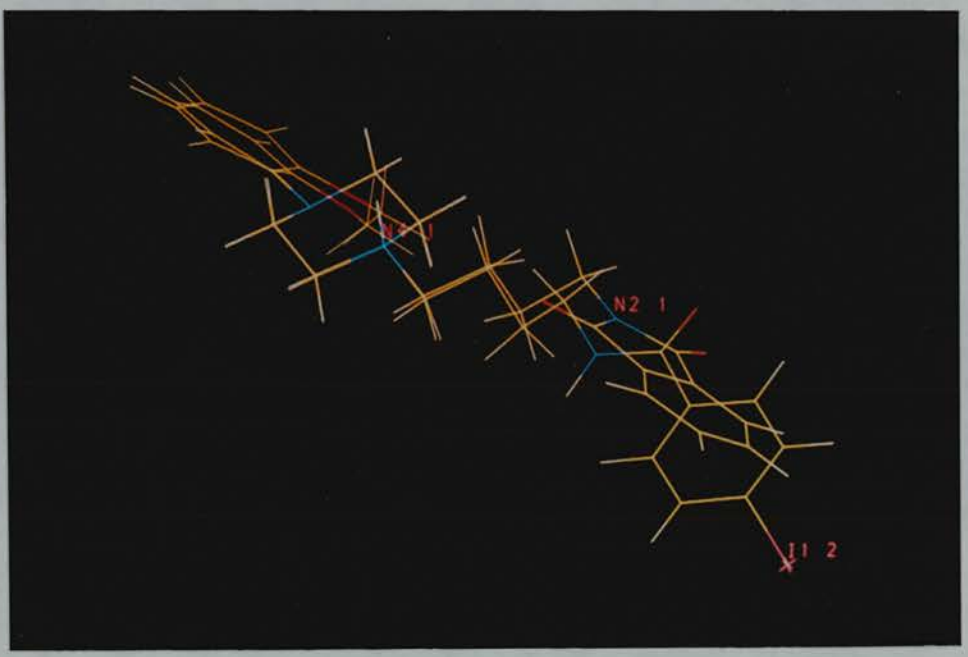


Figure 3–5: NAN-190 and IMD-1 superimposed.

ring are coplanar. The structure of IMD-1 has the benzamido ring rotated approximately 60° anticlockwise with respect to the methoxyphenyl ring when viewed from the iodine atom. This shift in the position of the benzamido group is responsible for a distortion of the butyl chain - atoms C2', C3', and C4' are moved with respect to the same atoms in NAN-190. It is interesting to note that the thermal parameters for these atoms are significantly higher in IMD-1 than NAN-190 (doubled in the case of C4').

3.2.5 Biological Significance of the Structures

The structures of three compounds similar to NAN-190 and IMD-1 have been solved by X-ray crystallography. They are 1-(2-methoxyphenyl)-4-[4-(2-phthalimido)but-2E-enyl]piperazine.HBr (NAN-190E) (Cabral, to be published) and the but-2E-ynyl derivative (NAN-190Y) (Muskett, 1992). These structures are identical to NAN-190 but have a double and triple bond respectively between C2' and C3' in the butyl chain. These three structures were superimposed manually using FRODO, using the piperazine ring as a common reference. The difference between the three structures is the length of the bond between C2' and

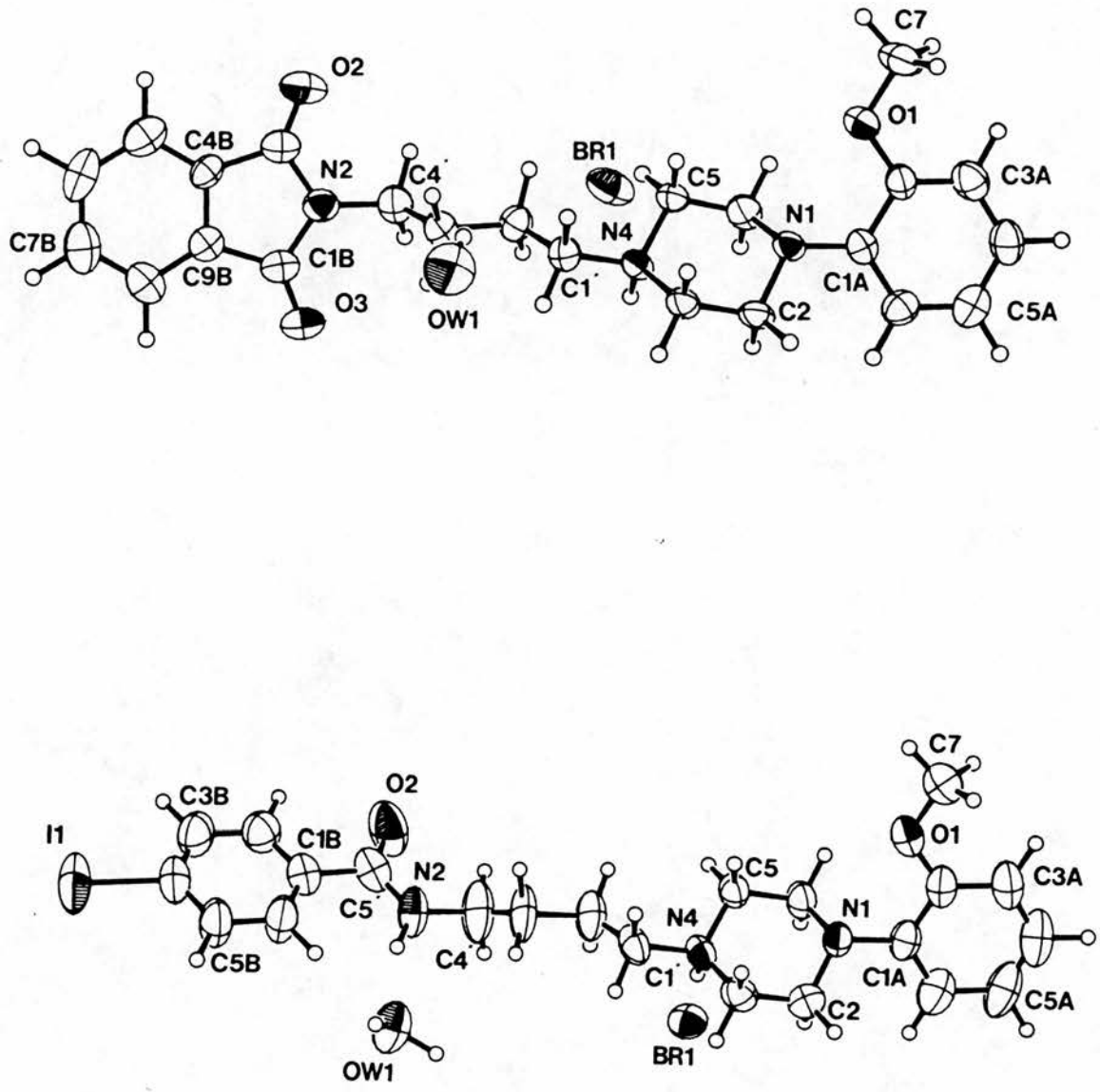


Figure 3–6: Structures of NAN-190 and IMD-1. Non-hydrogen atoms are shown as 50% thermal elipsoids

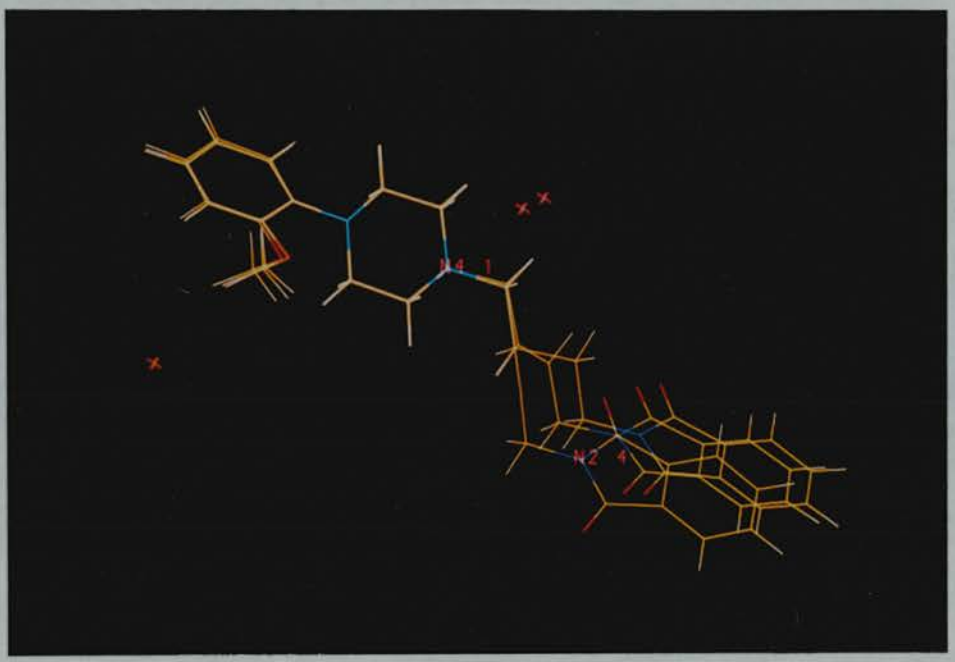


Figure 3–7: NAN-190, NAN-190E, and NAN-190Y superimposed

C3' (figure 3–7). The shortening of the bond results in a movement of the phthalimido group towards the piperazine ring. The phthalimido group in the but-2E-enyl derivative remains coplanar with the same group in NAN-190 but is shifted by 0.2 Å away from the plane of the group. The same group in the but-2E-ynyl derivative is shifted 1 Å and rotated approximately 10° anticlockwise looking from the piperazine ring. Distortion of the butyl chain is observed. This is due to the increased linearity of C1'-C2'-C3'-C4' atoms as the saturation of the C2'-C3' bond decreases.

The biological activity of two of these structures has been measured in rats *in vivo*. The inhibition constant (K_i) measures the antagonist effect of the compound against natural agonists for 5HT-1A receptors. The K_i for NAN-190 with the methoxy group replaced by a hydroxyl (NAN-190OH) or iodo moiety (NAN-190I) has also been measured (table 3–3). The results show that the nature of the group attached to the benzene ring at carbon C2A affects the antagonist activity. Highest activity is seen with a hydroxyl group substituted, lowest with an iodine. The known agonist serotonin, active *in vivo*, has a hydroxyl substituted on a benzene ring (figure 3–8), although not at an equivalent position to NAN-190. Comparision with other 5HT-1A agonists such

Molecule	K_i (nM)
NAN-190	1.5, 3
NAN-190E	6,15
NAN-190OH	1.5,2.5
NAN-190I	7.5

Table 3–3: Binding data for NAN-190 type compounds.

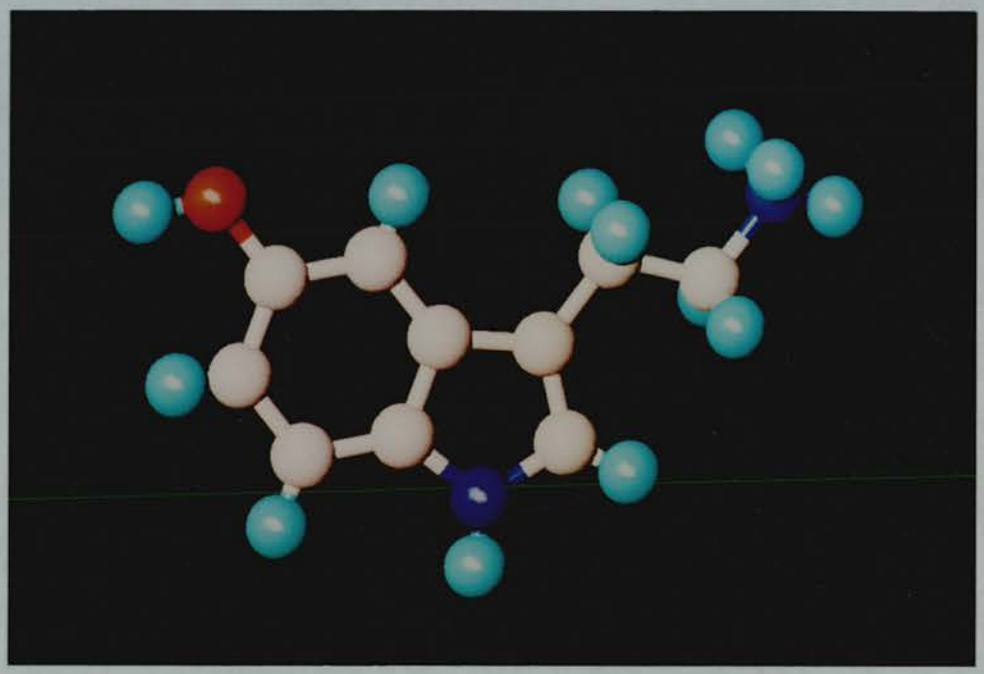


Figure 3–8: Structure of Serotonin.

as 8-OH-DPAT (figure 3–9) suggests that the important groups for activity are an aromatic ring substituted with an electronegative group, and a nitrogen atom some distance from this ring. In NAN-190 these groups are provided by the methoxybenzyl ring, and the nitrogen N4 of the piperazine ring. The antagonist activity of NAN-190 presumably derives from the butyl chain and phthalimido group. The difference in K_i for NAN-190 and NAN-190E suggest that either the position of the phthalimido group or the flexiblity of the butyl chain are the major determinant of antagonist activity. The binding data for the but-2E-ynyl derivative (NAN-190Y) would not clarify this situation as the butynyl chain is more rigid and linear, but the phthalimido group is shifted considerably (in the crystal structure).

These limited results allow some speculative conclusions to be drawn about the 5HT-1A receptor. The structure of serotonin, 8-OH-DPAT and NAN-190 suggest

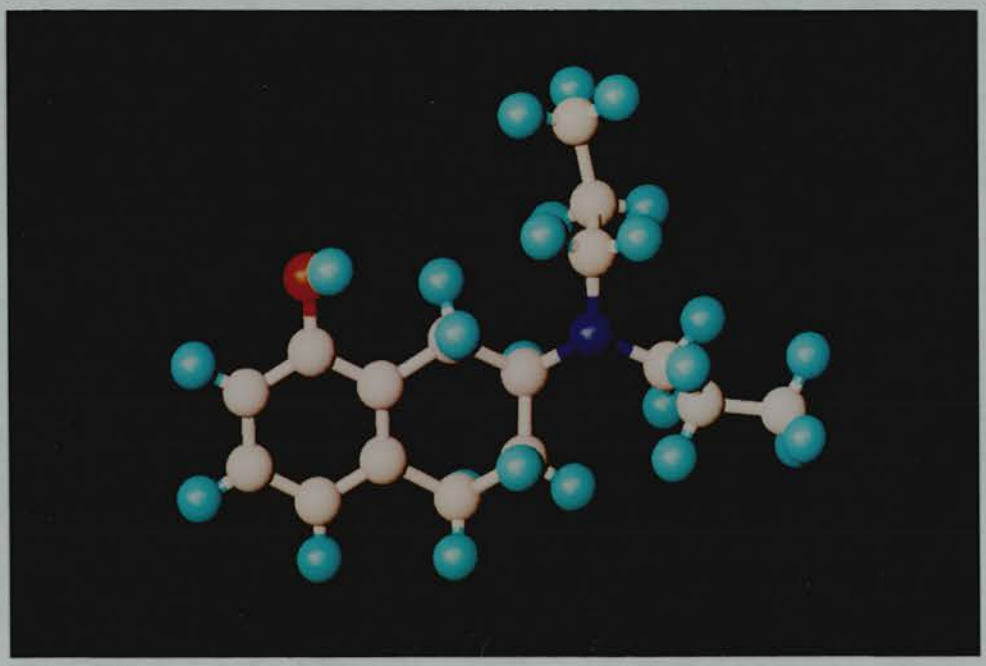


Figure 3–9: Structure of 8-hydroxy-2-(di-*n*-propylamine)tetralin.

that the receptor has a cavity which accepts a relatively small electronegative group, such as a hydroxyl. The cavity may be in a hydrophobic environment as this electronegative group is attached to a benzene ring. The receptor also has a group which interacts with a nitrogen atom on the ligand. This interaction is probably through a hydrogen bond, with the ligand nitrogen acting as a hydrogen bond donor. These three factors seem to be enough to activate to receptor. The antagonist NAN-190 has an aromatic group some distance from the nitrogen. This group is substituted with two oxygen atoms. The antagonist activity of NAN-190 may come from the interaction of this aromatic system with some hydrophobic part of the serotonin receptor. This may be stabilised by a hydrogen bond interaction with the electronegative oxygens. How the phthalimido group in NAN-190 diminishes 5HT-1A receptor activity remains unclear. It may prevent a conformational change in the receptor from occuring, or may interact with another site on the receptor to stabilise its binding to such a level that other more active ligands are unable to bind. It would be of interest to synthesise NAN-190 derivatives with the phthalimido group substituted with other related groups. This may help determine the nature of the interaction of this group with the 5HT-1A receptor.

3.3 Protein Crystallography

It can be seen that the solution of small molecule structures can be relatively straight forward. The determination of large molecule structures is less straight forward and is summarised in figure 3–10. The two major bottlenecks in macromolecular crystallography are obtaining large single crystals which diffract well, and then the determination of structure factor phases. The following section describes an attempt to obtain crystals of MUP. This is followed by preliminary X-ray diffraction studies on crystals of a2u grown from ammonium sulphate.

3.3.1 Purification and Crystallisation of Mouse Major Urinary Protein

Crucial to the crystallisation of a protein is its prior purification, preferably to homogeneity in both molecular mass and charge. Mouse major urinary protein has been previously purified using gel filtration followed by ion exchange chromatography (Finlayson *et al.*, 1968), and by gel filtration followed by preparative isoelectric focusing (Lorusso *et al.*, 1986). As an initial test, mouse urine from BALB/c male mice (courtesy of J. Bishop, Department of Genetics) was purified on the basis of molecular mass by gel filtration using the method of Finlayson (Finlayson *et al.*, 1968). The resulting sample was analysed using SDS-PAGE (figure 3–11) and 2-dimensional gel electrophoresis (figure 3–12). These techniques indicated the presence of a lower molecular weight protein and also at least six different charge variants of MUP. A revised protocol was used for purification: gel filtration, followed by chromatofocusing (equivalent to isoelectric focusing), and finally ion exchange chromatography. It was hoped that this protocol would purify MUP to a single molecular weight and charge species.

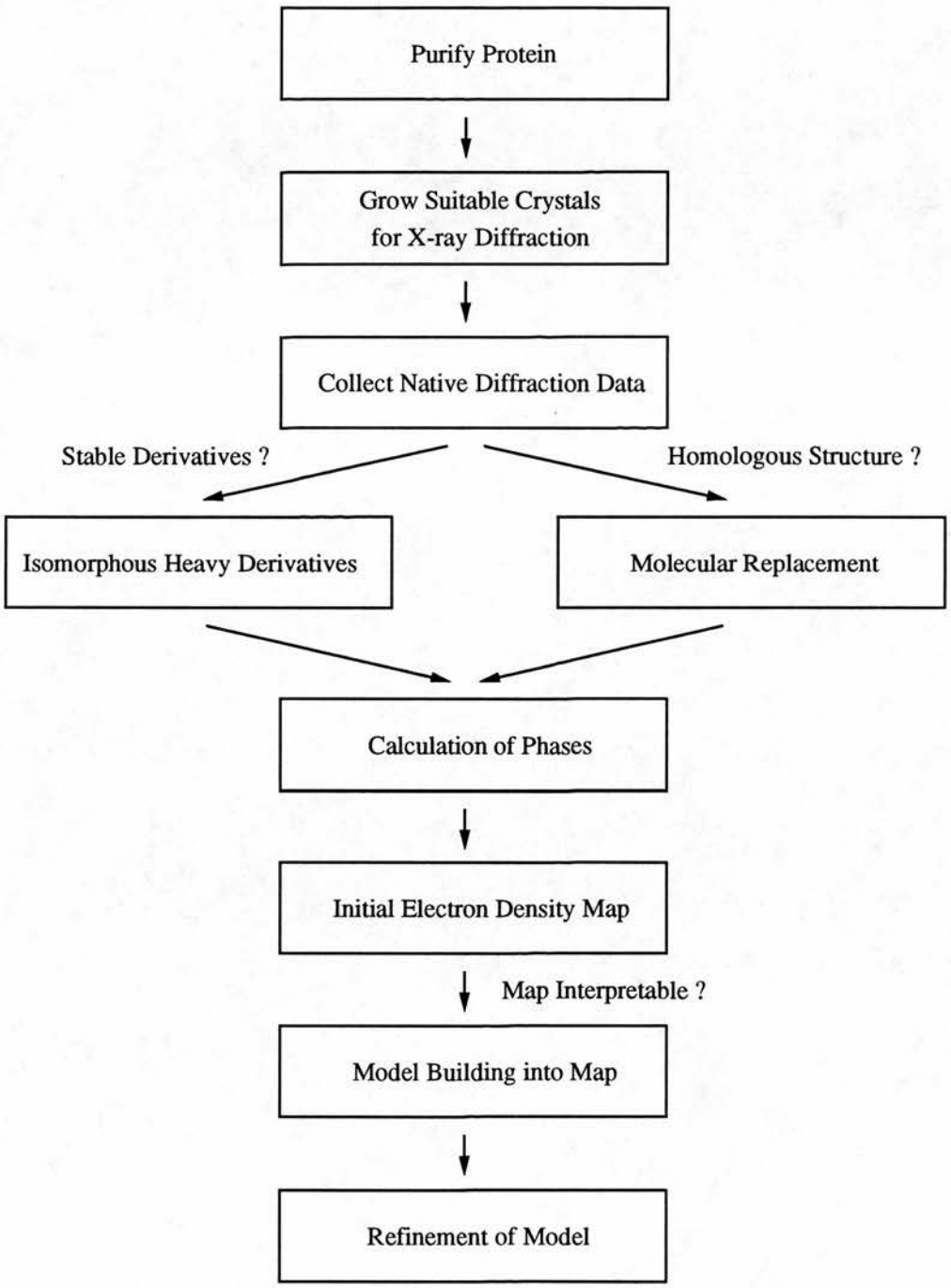


Figure 3–10: Generalised methodology for protein crystallography.

Figure 3–11: SDS-PAGE of MUP after purification by gel filtration using Sephadex G-100.

Collection and Preparation of Urine

Mouse urine was collected from male BALB/c mice housed in metabolic cages. After filtration through a sintered glass filter this urine was frozen and stored at -70° C. This urine was defrosted prior to purification, giving approximately 100 ml of dark brown liquid. The dark brown colouration was assumed to be due to oxidised catecholamines. The urine was centrifuged at 5000 rpm for 15 minutes to remove any large debris. The urine was then filtered through Whatman number 1 filter paper. This left approximately 70 ml of debris-free urine which was dialysed twice against 2 times 5 litres of phosphate-buffered (0.01M sodium phosphate, pH 7.5) saline (0.15M) (PBS) at 4° C. Dialysis tubing was D-9652, from Sigma, with a 12,000 dalton cutoff. Prior to use the tubing was boiled for 1 hour in EDTA (2 g/litre) then for 1 hour in deionized water. The dialysed urine (70 ml), which had lost much of its brown colouration, was concentrated using a YM10 filter in a 50 ml Amicon Ultrafiltration cell. This produced 15 ml of concentrated mouse urine which was stored at 4° C.

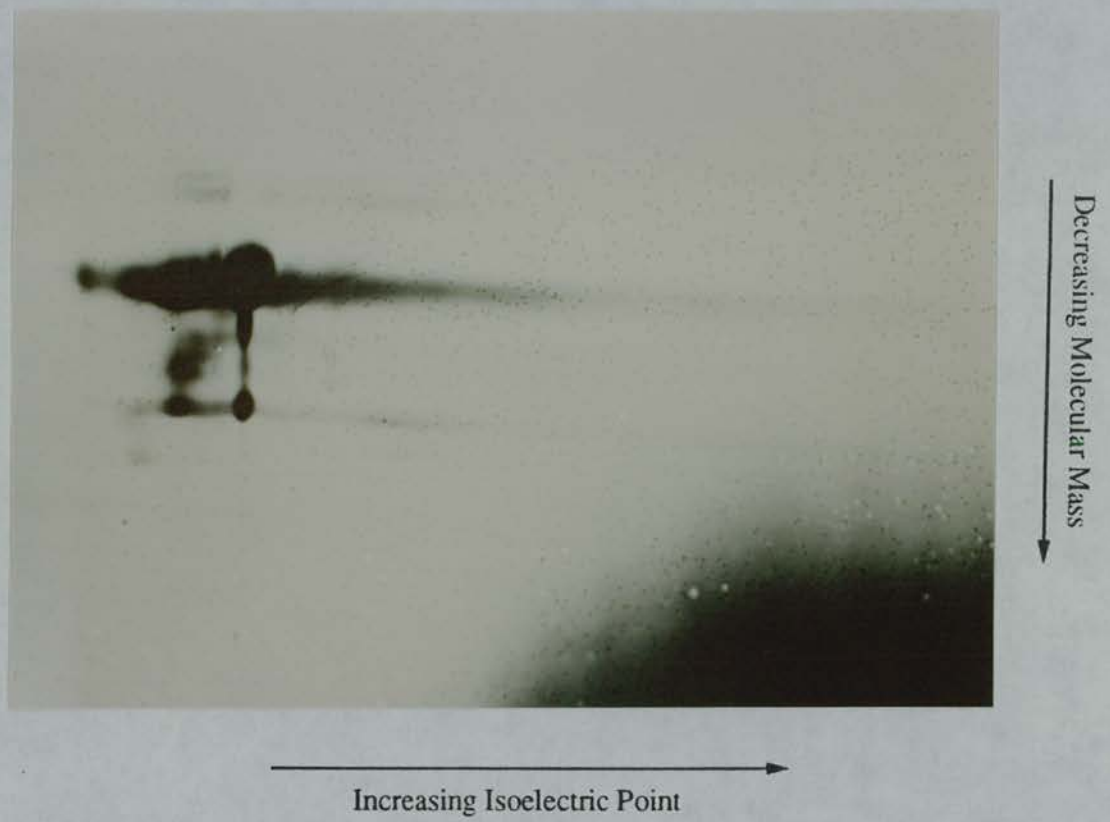


Figure 3–12: Two dimensional gel electrophoresis of MUP after purification by gel filtration using Sephadex G-100.

Peak	Fractions
I	1-30
II	40-54
III	61-63
IV	73-end

Table 3–4: Peaks pooled from gel filtration.

Gel Filtration

The gel filtration medium used was Sephacryl S-200HR, from Pharmacia. This was packed into a K26/70 column, from Pharmacia, using the Pharmacia Sephacryl S-200 packing instructions. The column was connected into a Pharmacia Fast Protein Liquid Chromatography (FPLC) system. The column was packed down using deionized water at a flow rate of 5 ml/min. The column was then inverted and washed with 2 bed volumes of PBS. The system was tested by running 1 ml of Blue Dextran (2 mg/ml) through the column at a flow rate of 2 ml/min. This produced a compact horizontal blue band which moved quickly along the column. The column was then cleaned thouroughly with PBS. The system was then used to purify 10 ml of the concentrated mouse urine. The sample was loaded onto the column and PBS passed through at 2 ml/min for 5 hours. During this process the initial dark brown band was seen to separate into 3 different bands; a lower light brown band, a darker brown middle band, and a yellow upper band. The absorbance at 280 nm of the column output was monitored (figure 3–13), and output collected as 2 ml fractions. This enabled the output fractions to be pooled into 4 peaks (table 3–4) The large peak II was assumed to be the MUP peak on the basis of previous purification work (Findlayson *et al.*, 1968). Peak II (approximately 60 ml) was dialysed twice against 2 times 4 litres of deionized water at 4° C overnight. The dialysed sample was then concentrated to 10 ml using a YM10 filter in an Amicon ultrafiltration cell.

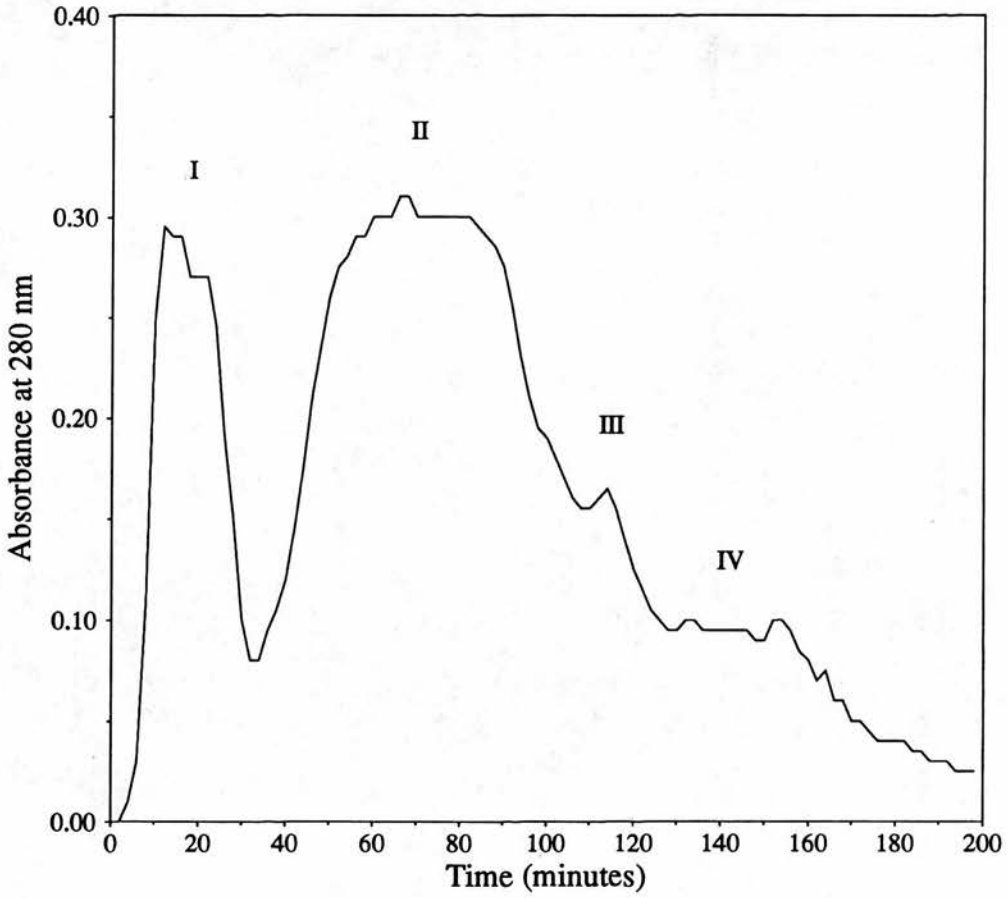


Figure 3–13: Progress of gel fitration of MUP using Sephacryl S-200HR. Protein concentration monitored by absorbance at 280 nm

Peak	Volume (ml)
a	4
b	7
c	30
d	9
e	8

Table 3–5: Pooled fractions from chromatofocusing.

Chromatofocusing

Chromatofocusing was carried out with a pre-prepared Mono-P HR 5/5 column, also from Pharmacia. This column was cleaned using 1 ml of NaCl (2M), then 1 ml of NaOH (2M), finally 1 ml of acetic acid (75%), with deionized water washes between each step. The eluent pH was monitored constantly using a flow-through pH electrode, the absorbance was monitored at 280 nm. The initial, high pH buffer was Bis-Tris (0.025M, pH 6.3). The elution buffer was Polybuffer 74, from Pharmacia, diluted 10-fold with deionized water, at a pH of 3. Samples of dialysed peak II (500 μl) were chromatofocused using FPLC. The initial buffer was a mixture of 75% Bis-Tris (buffer A) and 35% Polybuffer (buffer B) giving a starting pH of 5.5. The mixture was changed to 100% Polybuffer over 40 minutes. After each run the pH of the column was regenerated using 125 μl of NaOH (2M). This protocol resolved the sample into 2 main, large peaks with several smaller peaks (figure 3–14). From the constantly monitored pH trace it was possible to assign iso-electric points to these protein peaks (figure 3–14). This protocol was repeated 6 times with 500 μl samples of peak II and the fractions corresponding to the 5 different peaks pooled (table 3–5). All these samples were dialysed against 3 times 5 litres of deionized water at 4° C. Peak c was concentrated to 6 ml using a YM10 filter in an Amicon ultrafiltration cell.

Ion Exchange Chromatography

Ion exchange chromatography was with a pre-prepared Mono-Q HR 5/5 column, from Pharmacia. The base buffer (buffer A) was Bis-Tris (20mM, pH6.4), the elution buffer (buffer B) was the base buffer plus NaCl (0.35M). The

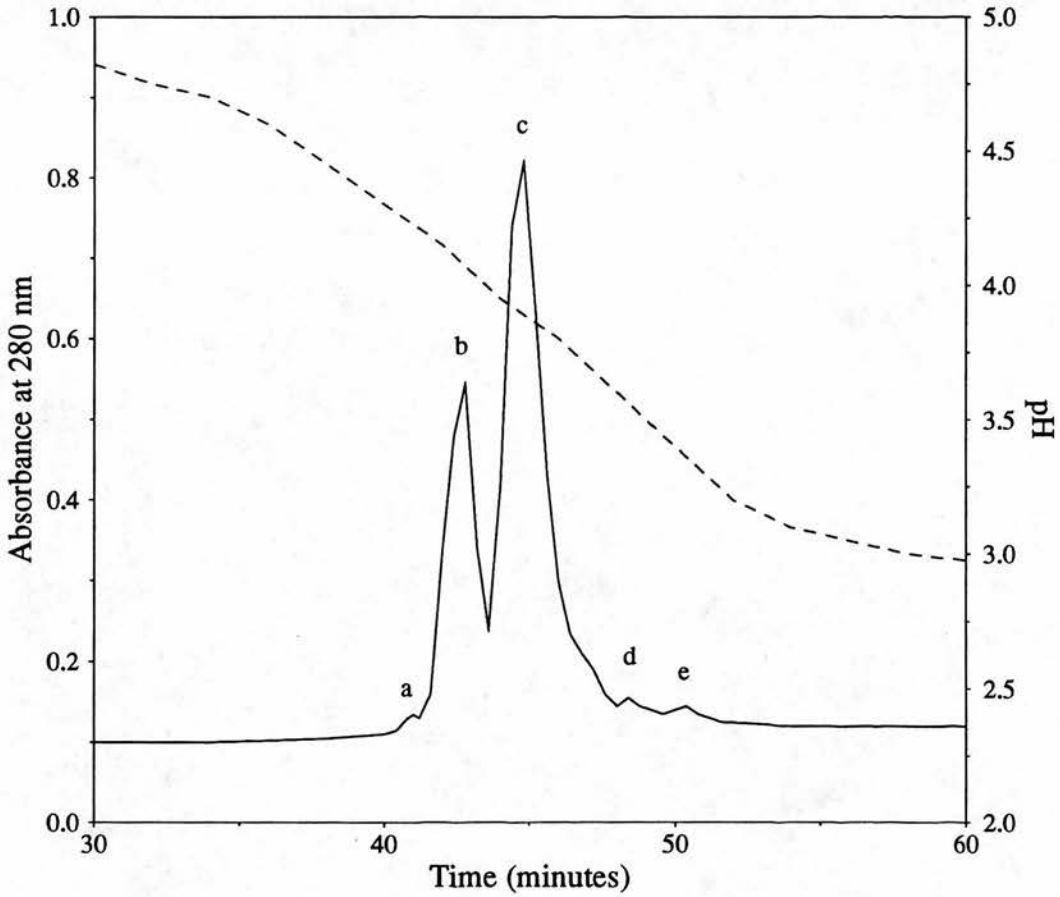


Figure 3–14: Progress of chromatofocusing of MUP (gel filtration peak II) using Poly-S. Protein concentration monitored by absorbance at 280 nm (solid line), pH monitored using continuous flow through pH-meter (dashed line).

Peak	Volume (ml)
1	18
2	10
3	9

Table 3–6: Pooled fractions from ion exchange chromatography.

Peak	Mass
a	100 μg
b	8.8 mg
c.1	28 mg
c.2	3.5 mg
c.3	1.2 mg
d	7.3 mg
e	<10 μg

Table 3–7: Mass of protein fractions from ion-exchange chromatography after freeze-drying.

concentration of elution buffer was varied as shown in figure 3–15. The column was washed with 100% elution buffer after each run to remove protein still bound. This method was repeated 8 times with 500 μl samples of peak c (figure 3–15). The fractions corresponding to the regions 1, 2 and 3 were pooled (table 3–6) These peaks were dialysed against 3 times 4 litres of deionized water at 4° C.

Analysis of Purified Protein

The peaks a, b, c.1, c.2, c.3, d, and e were freeze dried (table 3–7). Samples a and e contained only a very small dry mass of protein, therefore these samples were collected by washing the freeze drying vessel with 1 ml of deionized water.

SDS-PAGE Gel Electrophoresis

Gel electrophoresis used the method of Laemmli (Laemmli, 1970) with an LKB gel electrophoresis system (see figure 3–16). Staining of gels was with the Argonne protocol used for 2 dimensional gel electrophoresis (Andrews, 1986).

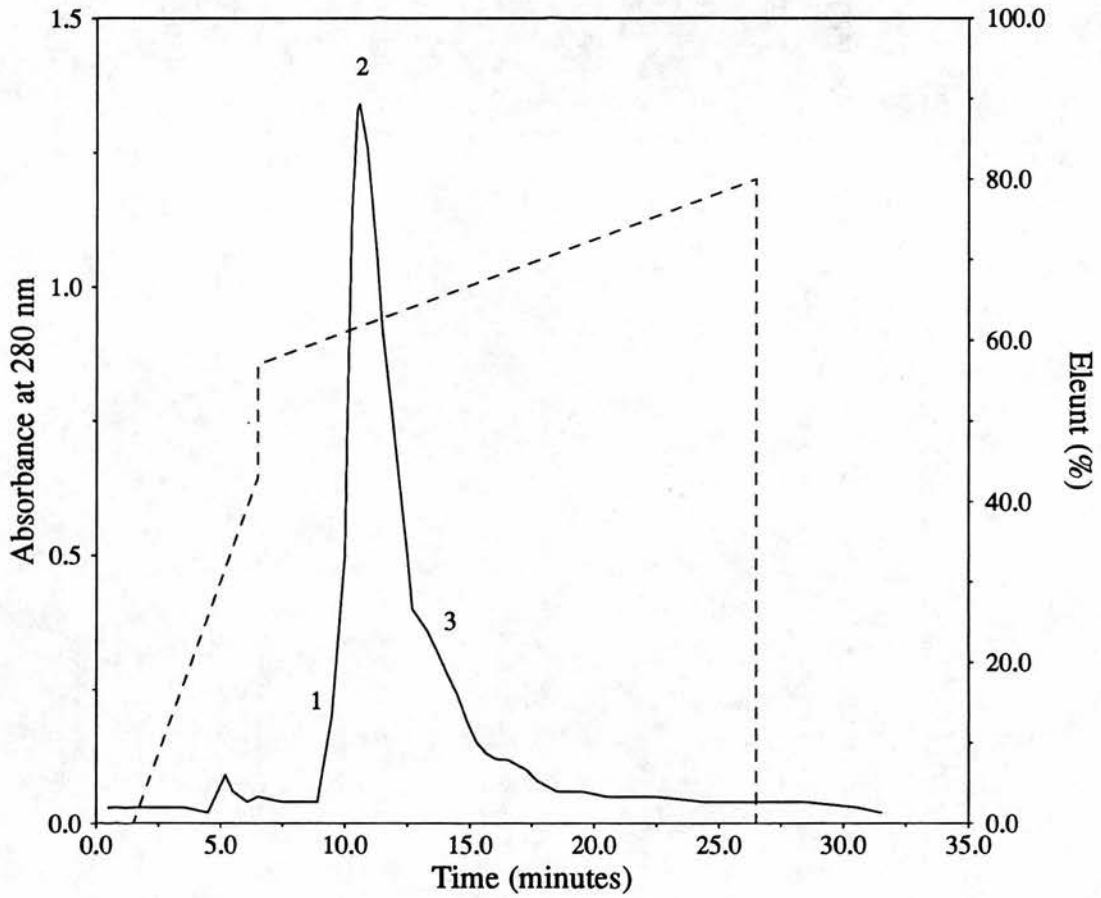


Figure 3–15: Progress of ion exchange chromatography of MUP using Mono-Q. Protein concentration monitored by absorbance at 280 nm (solid line), eluent concentration (dashed line).

Figure 3–16: SDS-PAGE analysis of purified MUP samples.

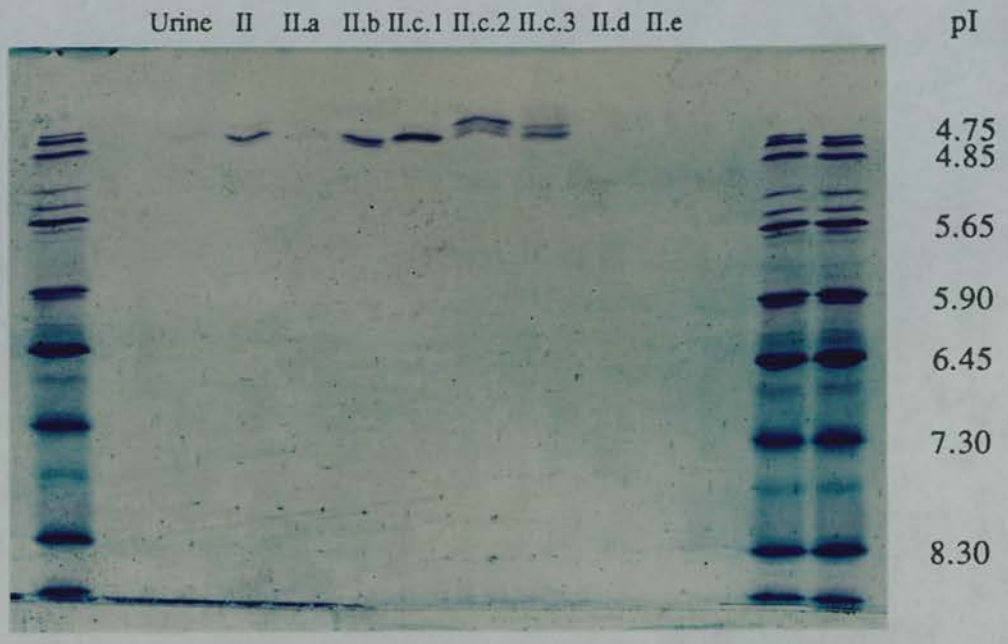


Figure 3–17: Isoelectric focusing analysis of purified MUP samples.

Two Dimensional Gel Electrophoresis

Two-dimensional gel electrophoresis used the method of Argonne with a custom built electrophoresis system at the ICI Central Toxicology Laboratory, Alderley Park. Staining of gels was with the Argonne protocol.

Isoelectric Focusing

Isoelectric focusing used pre-prepared Ampholine PAGplates (pH range 3.5 to 9), from LKB, with a horizontal electrophoresis tank from LKB. Staining of gels was with the Argonne protocol (see figure 3–17).

Protein Concentration

The protein concentration of samples was assayed using the method of Bradford (Bradford, 1976). Samples were assayed directly or diluted to give suitable absorbance reading at 595 nm, freeze dried samples were weighed out to give a concentration of 1 mg/ml when redissolved in distilled water. A standard protein concentration curve was determined for bovine serum albumin (BSA). The various protein samples were assayed twice with absorbance at 595 nm measured after 30 minutes and 2 hours. The protein concentrations derived from the standard curve were 10 fold lower than expected. It was assumed that this was due to the nature of the proteins in question, which possibly suggested that the binding of the assay dye was affected by the isoelectric point of the protein, the more acidic the protein the less dye binds. The standard protein BSA has an iso-electric point of 4.9, but the major isoelectric point of MUP, as determined during chromatofocusing, was 3.9. It is also possible that the difference in size of BSA (67 Kd) and MUP (18 Kd), coupled with the different surface charges may have determined the amount of dye that could bind. Therefore, a standard protein concentration curve was calculated using BLG, a readily available protein of similar size (18 Kd) to MUP but similar isoelectric point (5.2) to BSA. This standard curve is seen to over-estimate the concentration of the samples by at most a factor of two. The Bradford method is therefore not recommended for determining the concentration of MUP. Instead, the concentration of MUP samples was assessed visually from SDS gels stained with Coomassie Blue. The intensity of gel bands for MUP samples was compared to the band intensity for BLG samples of known concentration.

N-terminal Amino-acid Sequencing

In order to determine the identity of the purified protein the N-terminal amino-acid sequences were determined for samples II.b, II.c.1, II.c.2, II.c.3, and II.d (table 3–8). This sequencing was carried out by the ICI protein sequencing facility at Alderley Park.

Cycle	Sample					
	b	c.1	c.2	c.3	d	MUPI
1	?	E	E	?	SGA	E
2	E	E	E	?	?	E
3	A	A	A	A	FQA	A
4	S	S	S	S	(S)	S
5	S	S	S	S	?	S
6	T	T	T	T	FM	T
7	G	G	G	G	EG	G
8	R	R	R	R	RF	R
9	N	N	N	N	N	N
10	F	F	F	F	FG	F
11	N	N	N	N	NF	N
12	V	V	V	V	V	V
13	E	E	E	E	E	E
14	K	K	K	K	GK	K
15	I	I	I	I	I	I
16	N	N	N	N	N	N
17	G	G	G	G	GF	G
18	E	E	E	E	F	E
19	?	W	W	W	?	W
20	H	H	H	H	?	H
21	T	T	T	T		T
22	I	I	I	I		I
23	I	I	I	I		I
24	L	L	L	L		L
25	A	A	A	A		A
26	S	S	S	S		S
27	?	D	D	D		D
28	K	K	K	K		K
29	R	R	R	R		R

Table 3–8: Sequences identified by N-terminal sequencing of purified MUP samples. The sequence of MUPI is given for comparison. A single letter indicates an unambiguous assignment. Residues in brackets are tentative assignments. A question mark indicates that no residue could be assigned.

Crystallisation Trials

All crystallisation trials were carried out using the hanging drop method (McPherson, 1982). Coverslips were siliconised before use (with dimethyldichlorosilane) and wells sealed using petroleum grease. Freeze-dried protein was redissolved in distilled water to give 20 mg/ml for MUP. The crystallisation buffer was 50 mM citrate-phosphate unless otherwise stated. Well volumes of 1 ml were used with varying precipitant concentration and pH. Hanging drops were made with 5 μl of well buffer and 5 μl of protein solution. Various precipitants were used to try to crystallise MUP: NaCl, ethanol and AS. It was possible to obtain very small crystals with either ethanol (15addition of *n*-octyl glucoside (1mM) or dioxan (0.001%) to the crystallisation buffer did not improve crystal growth. The addition of PEG-4000 (1%) did seem to increase the likelihood of small crystals, but was not a significant improvement. Seeding experiments with small crystals grown from AS did not produce larger crystals. It was observed that the crystals were very sensitive to changes in temperature. Crystals grown at 10° C if moved to room temperature for as little as ten minutes would dissolve when returned to a constant 10° C. The crystals grown by AS precipitation regrew after first dissolving, crystals grown by ethanol precipitation remained partly dissolved. It was not possible to grow crystals large or stable enough for X-ray diffraction analysis using these conditions.

3.3.2 X-ray Diffraction Analysis of Alpha-2u-Globulin

Alpha-2u-globulin was purified from the urine of Fischer 344 rats by P. Phillips and E. Lock (ICI Central Toxicology Labs., Alderely Park). The freeze-dried protein was used for crystallisation trials.

Crystal Growth

As before all crystallisation trials were carried out using the hanging drop method. Coverslips were siliconised before use (using dimethyldichlorosilane) and

Figure 3–18: Single crystals of MUP grown from ammonium sulphate (10%) at pH 8.

wells sealed using petroleum grease. Freeze-dried protein was redissolved in distilled water to give 15 mg/ml for a2u. The crystallisation buffer was 50 mM citrate-phosphate unless otherwise stated. Well volumes of 1 ml were used with varying precipitant concentration and pH. Hanging drops were made with 5 μl of well buffer and 5 μl of protein solution. For a2u, crystallisation was observed over a wide range of pH and AS concentrations (table 3–9). The majority of these crystals were obviously twinned or very closely associated with each other (figure 3–19), although one well produced small single crystals (figure 3–20). Narrowing the pH and precipitant concentration range did not produce better crystals, instead larger twinned crystals were produced. The addition of *n*-octyl glucoside (1mM) to the crystallisation buffer did not improve crystal growth. It was not possible to grow crystals using NaCl as the precipitant. Using AS as the precipitant it was possible to obtain crystals large enough for X-ray diffraction analysis.

	AS Concentration (M)			
pH	0.5	1.0	2	3
3	xtal clump	xtal clump	ppt.	ppt.
4	ppt.	clumped ppt.	ppt.	ppt.
5	ppt.	large xtals	ppt.	ppt.
6	plate-like xtals	ppt.	xtal clump	xtal clump
7	small xtals	xtal clump	thin needles	ppt.
8	ppt.	ppt.	needles	large xtal

Table 3–9: Results of crystallisation trials for a2u.

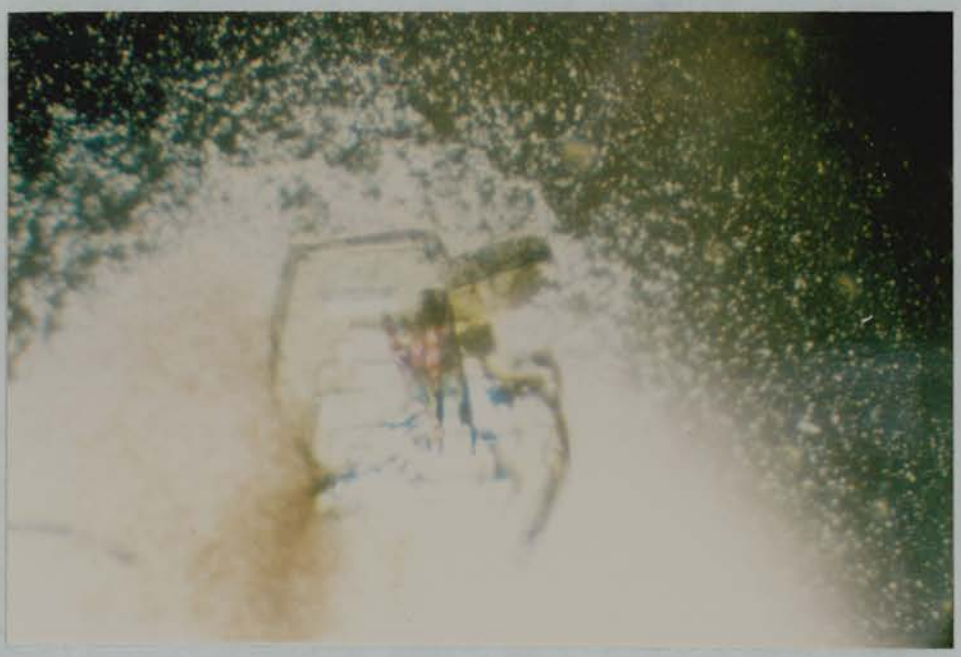


Figure 3–19: Twinned crystals of a2u grown from ammonium sulphate (1M) at pH 5.

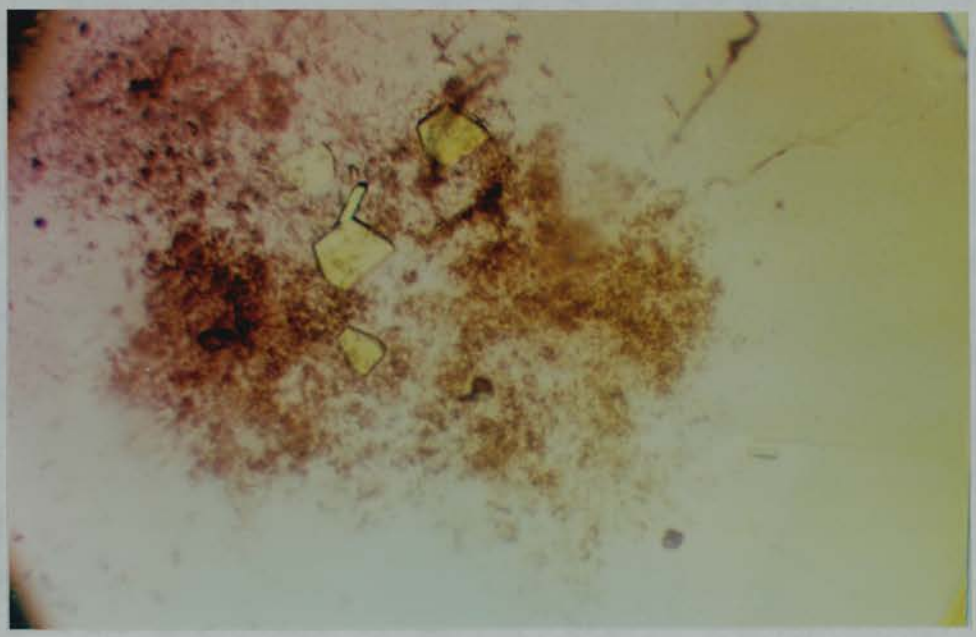


Figure 3–20: Single crystals of a2u grown from ammonium sulphate (3M) at pH 8.

Precession Photography

Crystals of a2u large enough for analysis were grown. An attempt to characterise the crystals using conventional radiation sources was made. A crystal (grown at pH 8.0, AS 3M) approximately 0.25 x 0.25 x 0.1 mm in size was mounted in a Lindemann glass capillary. This was mounted on a Supper precession camera against a sealed Cu tube (Phillips 1.5 kW). A still exposure for 1.5 hours gave some small spots (figure 3–21). It was not possible to centre the diffraction pattern on a major zone. The morphology of the crystal gave little information about the possible location of the principal crystallographic axes. It was concluded that either the zones observed were only minor, thus making centring difficult, or that the crystal was twinned. The crystal stopped diffracting after less than 24 hours in the X-ray beam. Another crystal (grown at pH 6.0, AS 2M) approximately 0.3 x 0.2 x 0.1 mm in size was similarly mounted in a Lindemann glass capillary. This was mounted on a Huber precession - rotation camera against an Elliot GX-20 rotating anode. A two hour still exposure showed spots which made it possible to centre the diffraction pattern. A 9° precession photograph exposed for 36 hours showed no spots because the crystal had

Figure 3–21: Still photograph of a2u crystal (see text for details).

stopped diffracting. The lifetime of the crystals on conventional sources made the use of synchrotron radiation neccessary.

Studies using Synchrotron Radiation

The synchrotron radiation facility at SERC Daresbury Laboratory, UK, was used. The fixed wavelength (1.4884 Å) source on beam-line 7.2 and the shorter wavelength (0.96 Å) wiggler on station 9.6 were used.

Film A crystal (grown at pH 6.0, AS 2M) approximately 0.2 x 0.2 x 0.1 mm in size was mounted in a Lindemann glass capillary. This was mounted on an Enraf-Nonius oscillation camera on beam-line 7.2 at Daresbury SRS. A red 1 spot collimator was used with a crystal to film distance of 100 mm. Still

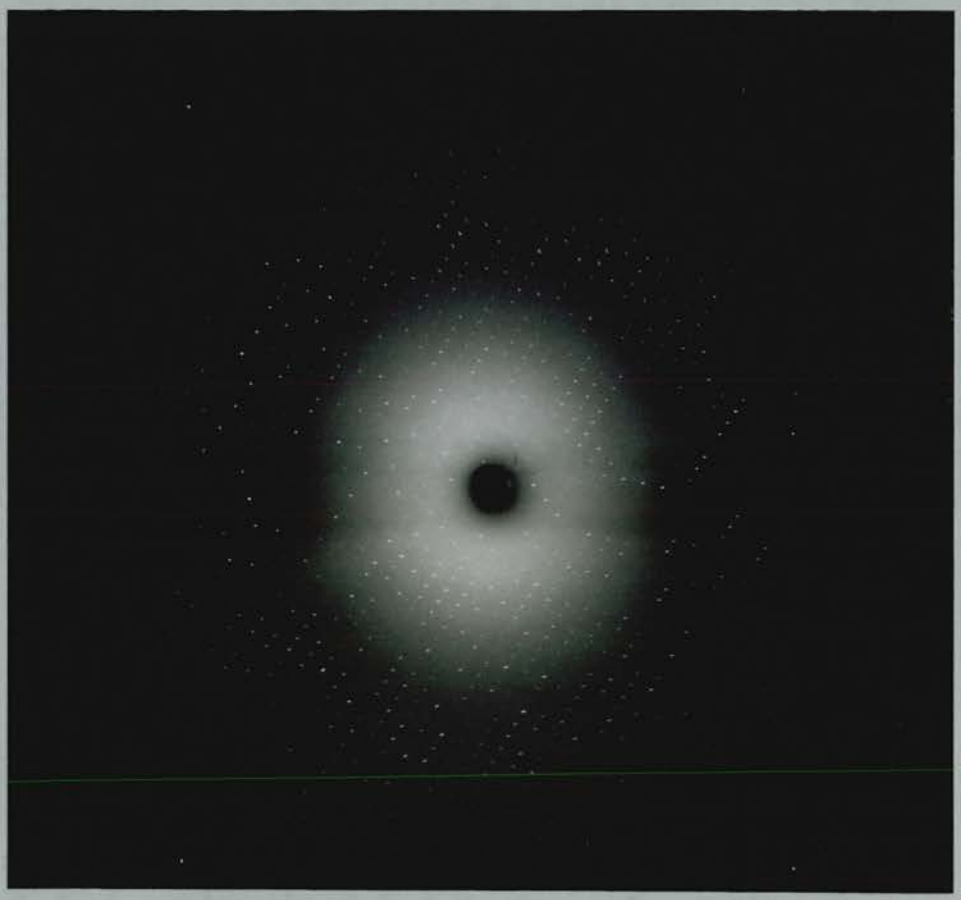


Figure 3–22: Rotation photograph of a2u taken at beam line 7.2, Daresbury SRS (see text for details).

exposures and rotation photographs (3° per exposure, 15 seconds/degree repeated 10 times) were taken at 0° and 90° (figure 3–22). Strongly diffracting spots were observed to 2.8 Å resolution. The rotation photographs suggested a monoclinic unit cell with $a = 56$ (Å), $b = 132$ (Å), $c = 119$ (Å), $\beta = 109.5°$, however this was considered unreliable because of possible twinning of the crystal. Rotation photographs were taken between 90° and 111°, at which point the crystal stopped diffracting. Two further attempts were made to characterise a2u crystals using beam-line 7.2 but both crystals proved to be twinned.

FAST Area Detector A crystal (grown at pH 4.0, AS 2M) approximately 0.25 x 0.05 x 0.05 mm in size was mounted in a Lindemann glass capillary. This was mounted on the FAST area detector system at beam-line 9.6 at Daresbury

SRS. Diffraction was seen out to 3 Å resolution with a crystal to detector distance of 80 mm and a detector swing angle of 0°. These still images were later merged with the program FASTMERG to reconstruct a rotation image using the program FASTPS (figure 3–23). Reflections were found from 3° oscillation ranges (30 images of 0.1° oscillation, 5 seconds per image) at 0° and 90°. These reflections were autoindexed using the AUTI subprogram in MADNES (Pflugrath and Messerschmidt, 1988). However, it was not possible to refine these parameters in the REFINE subprogram of MADNES. It was concluded that the crystal was twinned.

3.3.3 Discussion

The crystallisation of a2u and MUP has been reported elsewhere (Böcskei *et al.*, 1991). Crystals of a2u were grown from PEG 3350 by the hanging drop method. The protein was purified from male Wistar rats using gel filtration only (through Sephadex G-50). No indication is given of the purity of the protein samples used in crystallisation trials. It is unlikely that a single charge species was produced using gel-filtration alone. However, different strains of rat typically produce different levels of a2u varients. It is possible that the Wistar strain used produces urine rich in one particular charge varient of a2u.

The purified a2u from ICI used in the crystal trials described previously was not purified to homogeneity - as assessed by two-dimensional electrophoresis. The charge heterogeneity observed could have accounted for the problems encountered in crystallisation. That protein purity is a critical factor in obtaining reproducible crystal growth has been reported elsewhere (McPherson, 1982). Purification of protein samples to one single charge species resulted in reproducible and improved crystal growth for monoclonal Fab fragments (Orbell *et al.*, 1988). However, crystals of a2u suitable for diffraction analysis were obtained from a relatively crude purification scheme, suggesting that either the choice of precipitant or rat strain is of importance. The Fischer 334 rat strain produces a broad range of a2u charge variants - as indicated by two-dimensional

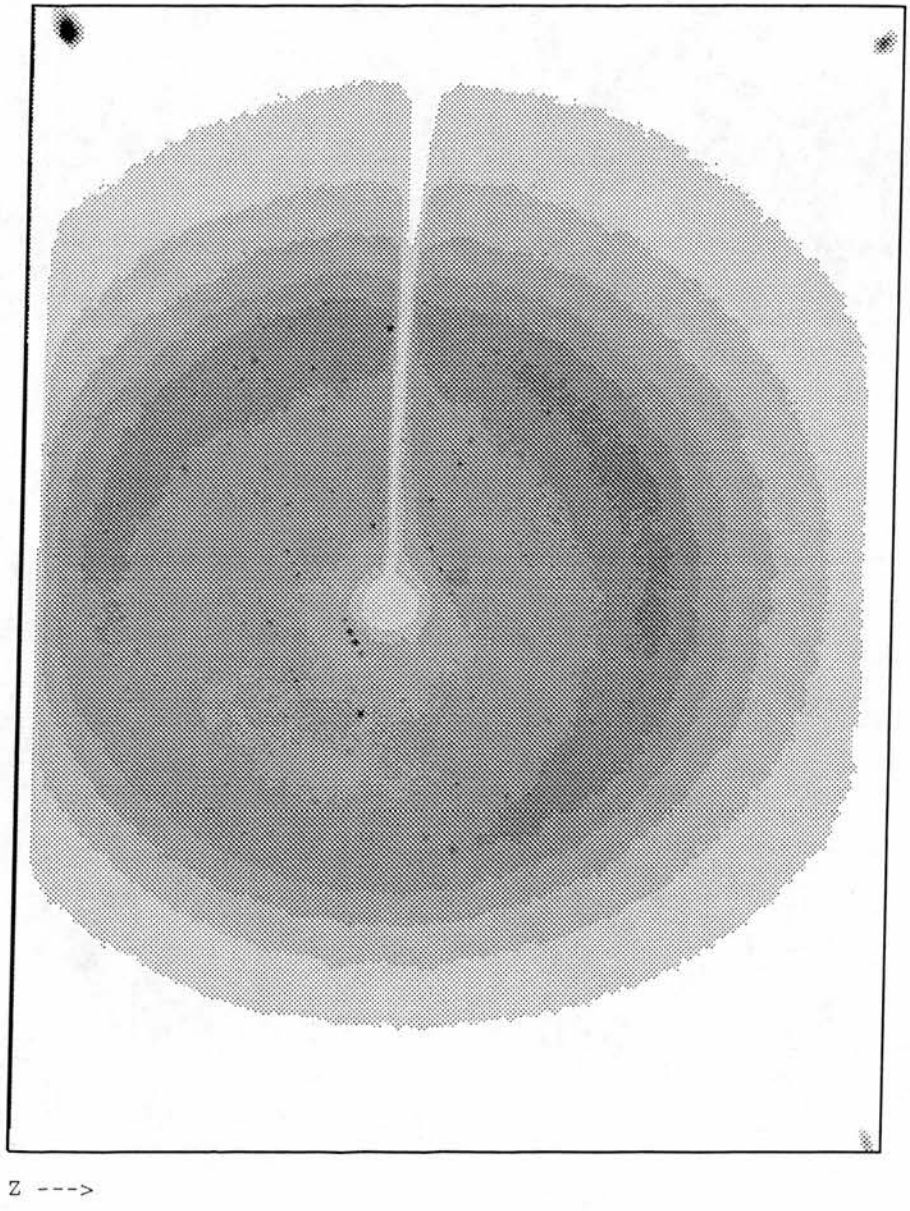


Figure 3–23: Rotation image for a2u reconstructed from still images taken using the FAST detector system at beam line 9.6, Daresury SRS (see text for details).

electrophoresis of purified samples. It is possible that suitable crystals of a2u could be obtained from ammonium sulphate if the protein sample were purified to one charge species.

The purified MUP used in the work here was less heterogeous in charge than the a2u. The results of isoelectric focusing indicate that sample II.c.1 had been purified to a single charge species (figure 3–17). Use of this fraction in crystallisation did result in crystals (as described above). These crystals were not large but showed a regular morphology, unlike the a2u crystals grown from ammonium sulphate. However, these crystals of MUP were very unstable when moved from 10° C to room temperature. The crystallisation of MUP by other workers, produced large crystals (0.5 x 0.3 x 0.3 mm) which were stable. It is noted that these crystals were obtained with cadmium ions ($CdCl_2$) as the precipitant. As with a2u a different strain of animal was used - in this case Tuck No. 1 white mice rather than BALB/c mice. The purification of MUP from Tuck No. 1 mice used the same procedure as a2u from Wistar rats. A single charge species of MUP is therefore unlikely with this technique, unless that particular strain of mouse produces predominantly one MUP variant. The crystal structure of MUP showed Cd^{2+} ions bound at the surface of the protein. The crystallisation of RBP used both Cd^{2+} and PEG 6000 as precipitants (Newcomer *et al.*, 1984). BLG has also been crystallised using Cd^{2+} as the precipitant (Green and Aschaffenburg, 1959). It seems possible that Cd^{2+} acts as a stabilising agent in the crystallisation of MUP - the four sites on one molecule link to four neighbouring subunits in a distorted tetrahedral manner. The binding of the Cd^{2+} ions is mainly by carboxylic groups and histidines. MUP has a range of isoelectric point variants but all are acidic - with pIs of 4.0 and below. The acidic nature of MUP can be rationalised by the high number of glutamate and aspartate residues in the amino acid sequence. It is possible that addition of Cd^{2+} to the crystallisation mixture may promote the growth of bigger, stable crystals. Ideally the amino acid sequence of the isoform used for crystallisation trials should be known. It is also possible that a surface amino acid could

present an unfavourable crystal contact in one isoform and not another, the latter therefore crystallising much more readily.

Chapter 4

Molecular Modelling

4.1 Protein Molecular Modelling

Knowledge of the structure of macromolecules at the atomic level is necessary to understand fully their activity and function. This applies to proteins, poly-nucleotides such as RNA and DNA, and also carbohydrate polymers. However, only proteins will be considered in this discussion as they, to date, have been best studied structurally. As can be seen from any biochemistry text book, proteins are involved in most areas of biochemical interest in living organisms. They provide much of the machinery for DNA replication and expression - to produce other proteins. They provide the catalytic enzymes for metabolic pathways. Proteins can also be purely structural as in the extracellular matrix. Immunoglobulins, a major component of the immune response in higher animals, are proteins. The cell surface receptors which are involved in cellular regulation are proteins, as are many of the hormones and growth factors which bind to these receptors. An understanding of how a particular enzyme catalyses a reaction and what determines its substrate specificity can only be obtained from the atomic structure of the protein in question (Pincus and Scheraga, 1981). How antibodies bind to their antigens with such specificity has only recently begun to be understood as antigen/antibody complexes have been solved by X-ray crystallography (Tello *et al.*, 1990). The problem of designing drugs to specific targets can be made easier if the structure of the target is known. Unfortunately, the rate at which protein structures can be determined is still slow and the process usually complex. Advances in NMR techniques have made

it possible to determine the structure of small proteins up to 150 amino acids in size (Clore et al., 1990). The study of larger proteins still requires the use of X-ray crystallography which, despite advances in X-ray sources and X-ray diffraction detection equipment, still remains a slow process. Therefore, much effort has gone into the modelling of a protein's structure from other, previously determined, structures. This chapter presents results for the modelling of a2u which was carried out in an attempt to understand the protein's ligand binding activity. First an introduction to some of the techniques involved in this modelling is given.

4.1.1 Why Molecular Modelling is Possible

The number of protein sequences available is far greater than the number of protein structures, solved either by X-ray crystallography or NMR. At present there are approximately 17,000 protein sequences deposited in the GenBank sequence database (release 71.0), derived from both protein and DNA sequencing. However, unique atomic coordinates for less than 500 hundred of these sequences have been deposited in the Brookhaven Protein Databank (PDB; release 59). The task would seem hopeless: to determine the structures of a rapidly increasing pool of protein sequences by either crystallography or NMR but this problem is simplified by the observation that the protein structures determined so far show a limited number of structural motifs. The same secondary structure motifs, α-helix, β-sheet, and hairpin turns occur in proteins with very different sequences. These secondary structure units assemble to form common tertiary structures. Often these tertiary structures are associated with specific biological functions, such as nucleotide binding. Alternatively, a tertiary fold may be recognised in many different proteins because of an evolutionary link, for example $\beta\alpha\beta$ barrels (Lesk *et al.*, 1989). The structures determined so far indicate that there is information about a protein's structure in both its amino acid sequence and its function. The problem therefore becomes one of extracting information from a sequence in order to model our closest approximation to the structure of that sequence. That the protein sequence

determines tertiary structure is a well established concept, merely from the observation that proteins can be unfolded and refolded to the same active state (Anfinsen, 1973). The aim is to determine the structure of this active state purely from the sequence - the folding problem is a massive task. To make this task even possible to contemplate, assumptions have to be made. It is assumed that similarity in sequence implies similarity in structure. This is not always the case, as short peptides can often form more than one stable secondary structure (Argos, 1987). For the protein structures already solved, local sequence similarity does not imply a structural similarity when there is neither an evolutionary nor functional explanation to support this (Sternberg and Islam, 1990). Conversely, it is assumed that similarity in function implies a similarity in structure. Again this is not neccessarily the case, as shown by the helix-turn-helix and zinc finger structures, both of which bind DNA (Härd *et al.*, 1990; Omichinski *et al.*, 1990). The two major assumptions are flawed unless taken together. At the present time, we can only be confident about molecular modelling when we study sequences which show both functional and sequence similarity. This requires that the protein's amino acid sequence and its function be known. Historically, protein sequences were derived directly from purified protein using chemical and enzymatic methods. This required that the protein be available in large quantities, which in turn meant it was often well-characterized biochemically. More recent DNA sequencing methods make it possible to determine the amino acid sequence of a completely unknown protein. Therefore, a large number of protein sequences are known without the corresponding functional information being available. Often functional status is implied from sequence homology with proteins of known function, an assumption beset with pitfalls given what has just been said. However, there are still many sequences available for proteins of known biological function for which models can be constructed. The interactions, both covalent and non-covalent, between residues which determine tertiary structure are understood. Therefore, in theory it should be possible to simulate mathematically the folding of a protein in order to determine the tertiary structure. Practically, however, this is not possible due to the massive number of calculations involved. Instead, to model tertiary

structures the assumptions outlined above have to be used in conjunction with the structures already determined and supplemented by biochemical evidence.

4.1.2 Homology Modelling

Tertiary structure cannot yet be determined *ab initio* from primary structure using energy calculations. An alternative approach is to determine the rules that govern tertiary structure empirically, from analysis of crystal structures. We are looking for the complex function that relates sequence and structure:

$$x = f(sequence) \tag{4.1}$$

This function transforms a one dimensional vector (the sequence) to a three dimensional result (the structure), and as a consequence is very complex. Our knowledge about the empirical relationship between sequence and structure is still limited, a point which must be borne in mind. It is possible to attempt to derive empirical rules because of the observation that protein sequences and structures fall into families. That is to say, each protein sequence does not code for an absolutely unique tertiary structure: there are proteins with similar sequences that form similar tertiary folds. The existence of structural families of proteins is a result of divergent evolution, while functional families result from both divergent and convergent evolution. The chemical nature of the genetic code introduces spontaneous point mutations over time in genes which encode for proteins. Mutations may be lethal and obviously will be selected against. Alternatively, mutations can be neutral having no effect on protein activity or organism viability, such changes will be neither selected for nor against. Between these two extremes there are mutations that effect protein viability and/or protein activity/function. These will be selected for or against depending on what is most biologically appropriate. For example, a single gene coding for a protein existing many millenia ago can have been passed to many different organisms by evolutionary species divergence. The mutations of these subsequent genes will have produced different protein sequences for what is still the same protein in each organism. Obviously the closer the organisms are evolutionarily

the more similar the sequences will remain, although the mutation rate varies between different organisms. As time proceeds and the number of mutations increases the similarity of sequences is obscured by neutral mutations which do not effect the structure or function. However, those residues which are vital for a particular function will remain conserved. It is also possible to detect protein sequences which have lost these vital residues, with a consequent change in function, whilst still retaining enough sequence similarity to be identified as having the same common ancestor. In other cases functionality is retained by conservative mutation of vital residues. In the case of convergent evolution, two different proteins, possibly with different functions evolve to the same function. This may be by the occurrence of the same vital residues in both, or alternatively a different distribution of residues both carrying out the same function. Although biochemical evidence may suggest two proteins are related the sequences and therefore probably the structures are not. For example subtilisin and tryspin are both proteases with catalytic serine residues but no other structural similarities (Garavito *et al.*, 1977). The comparison of protein sequences is useful in determining evolutionary relationships between organisms. When trying to establish empirical relationships between sequence and structure a reasonably large number of protein structures needs to be considered. Fortunately, X-ray crystallography and NMR have solved approximately 400 unique structures to date. It is therefore possible to study the relationship between families of divergently related proteins with both similar and non-similar functions. It is possible to use this information to model the structure of some proteins from their sequence (Sander and Schneider, 1990). A brief guide to the comparison of protein sequences and its application to homology modelling is presented next.

Comparison of Sequences

The comparison of amino acid sequences is based upon the pairwise comparison of sequences. The common approach is to construct a matrix of size nm, where n and m are the lengths of sequence 1 and 2 respectively. Each matrix element can be assigned a value based on a residue exchange scoring system. The simplest

scoring would place a one where amino acids are identical and a zero where they are not. In practice more sophisticated matrices are used which may rely on conservation of the physical properties of residues. The scoring matrix often used is the Dayhoff point accepted mutation (PAM) matrix. This is derived from observations of preferred and avoided substitution frequencies in sequence alignments over 71 different protein families. This matrix reflects the biological pressures acting upon residue mutation which need not be directly related to the chemical properties of amino acids.

Dynamic Programming

Once the comparison matrix has been generated the problem is to determine the best alignment path through this matrix. Dynamic programming methods have been developed which transform this original comparison matrix. The algorithm considers all possible alignments or matrix paths by keeping the maximum running sum of the residue exchange values at any matrix point. The best path through the transformed matrix can be found by starting at the lower right hand corner and following the highest values, provided that a move is neither made to the right or downwards. The transformation of the matrix can be weighted by penalty scores for gap initiation and gap extension (Needleman and Wunsch, 1970).

Multiple Sequence Alignment

The existence of protein families implies the existance of several related sequences, how can we align more than a pair of sequences at a time? The problem can be broken down into three different possible tasks:

- Aligning several sequences simultaneously
- Aligning a new sequence to several sequences already aligned
- Aligning two sets of multiple alignments to each other

To align several sequences simultaneously a multi-dimensional implementation of the Needleman-Wunsch algorithm could be used (an N dimensional comparison matrix for N sequences). Unfortunately, the computational cost limits the number of sequences to 3 for a full matrix analysis and 8 for a reduced analysis. Alternatively, the sequences are aligned to one another as pairs and these pairwise alignments analysed to produce the multiple alignment (Higgins *et al.*, 1989; Vingron and Argos, 1990). This method is commonly used as the computational demands rise proportionally to N^2 rather than to the power of N.

Sequence Alignment and Tertiary Structure

How can we best use the results of these sequence alignment techniques to derive the structure of a protein? Usually, the closeness of two sequences is expressed in terms of sequence identity and similarity. Sequence identity is the number of identical amino acid matches between the two sequences, often expressed as a percentage. Sequence similarity is a score based on the number of identical and non-identical but conservative matches between the two sequences. The sequence identity score can be an indicator of structural identity provided functional similarities are apparent. Studies on crystal structures indicate that sequence identities above 50% for functionally related proteins strongly suggest very similar tertiary structures. At sequence identities of 30% the confidence in structural similarity even for functionally related proteins has dropped rapidly. The match between aligned sequences and aligned three dimensional structure has been assessed for the known structures. A relationship between sequence identity, structural identity and alignment length is suggested (figure 4–1; Sander and Schneider, 1990). Although the structures may be similar at low levels of sequence identity the problem is knowing whether the alignment relates the sequence and structural similarity correctly. The question becomes: which parts of the alignment are reliable and which must be treated with some suspicion? Computational methods are being developed which can be used to help determine the reliability of different regions of a sequence alignment (Vingron and Argos, 1990).

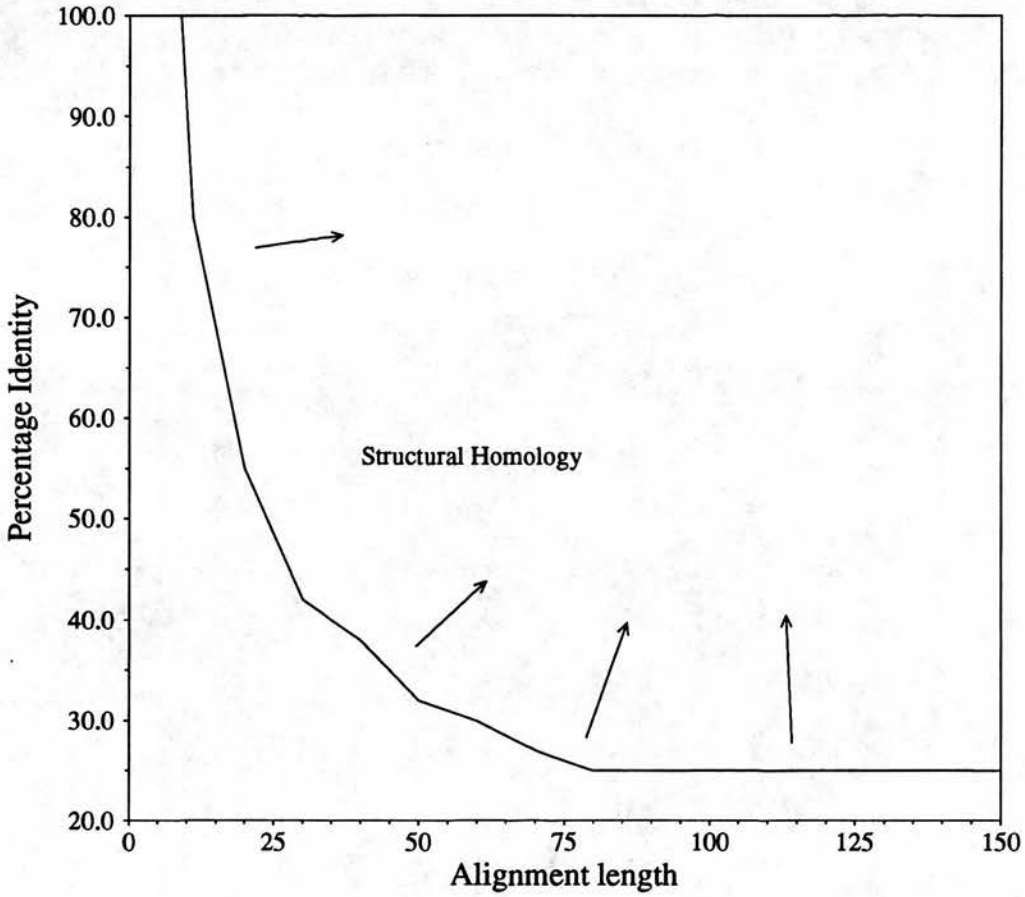


Figure 4–1: Plot of percentage sequence identity required for structural similarity at different sequence lengths.

The Modelling Template

The techniques are available to align the sequence of a known structure to the sequence of an unknown structure. If the alignment indicated a sequence identity of more than 70% it would be possible to mutate the residues of the known structure to those of the new sequence. This model would not be expected to differ greatly from the crystal structure. Alternatively, sequence identity levels of 30% with several structures may require information to be combined from all of them to generate a core conserved region around which the rest of the structure can be built. Further, at such low levels of sequence identity problems arise in successfully predicting the conformation of the loop regions. Practical expamples of modelling are given by: Blundell *et al.*, 1987; Greer, 1990; Havel and Snow, 1991. The techniques used for modelling at low levels of sequence identity will be discussed with practical examples later on in this chapter. First a brief background to the organisation of folded proteins and the forces that stabilise them is given. The following section deals with the factors that determine protein structure.

4.2 Protein Structure

The crystallographic determination of protein structures has consistantly shown the presence of distinct levels of organisation in protein structure (Chothia, 1984; Richardson, 1981). These levels of organisation are only an empirical description of protein structure and probably do not represent the way proteins fold to form their native structure. The question of protein folding will not be addressed here due to its complexity. Instead, a brief description of the protein structural heirachy observed from crystal structures is given. This is followed by a description of the forces between atoms that stabilise protein structure.

4.2.1 Hierarchy of Protein Structure

Analysis of protein structures can be rationalized in terms of a hierarchyof organisation; the sequence of amino acids, local association of these amino acids to form helices or extended strands, association of these units to form independent domains, association of these domains to form the tertiary structure, and in some cases the association of tertiary units.

Primary Structure (the Sequence)

Proteins are polypeptide chains, that is amino acid residues linked by amide (peptide) bonds. In nearly all organisms there are only 20 amino acids prescribed by the genetic code, all of which except glycine are the L-isomer. The peptide bond between amino acid residues is constrained to either the *trans* or *cis* form, the *trans* form is much more dominant in protein structures solved so far. A significant energy input is required to rotate this peptide bond, thus making the backbone of the polypeptide chain less flexible. All proteins have this same backbone connectivity, different folds for different sequences are a consequence of the nature of side chain groups for each residue which is determined by the amino acids sequence. The 20 amino acids can be grouped in

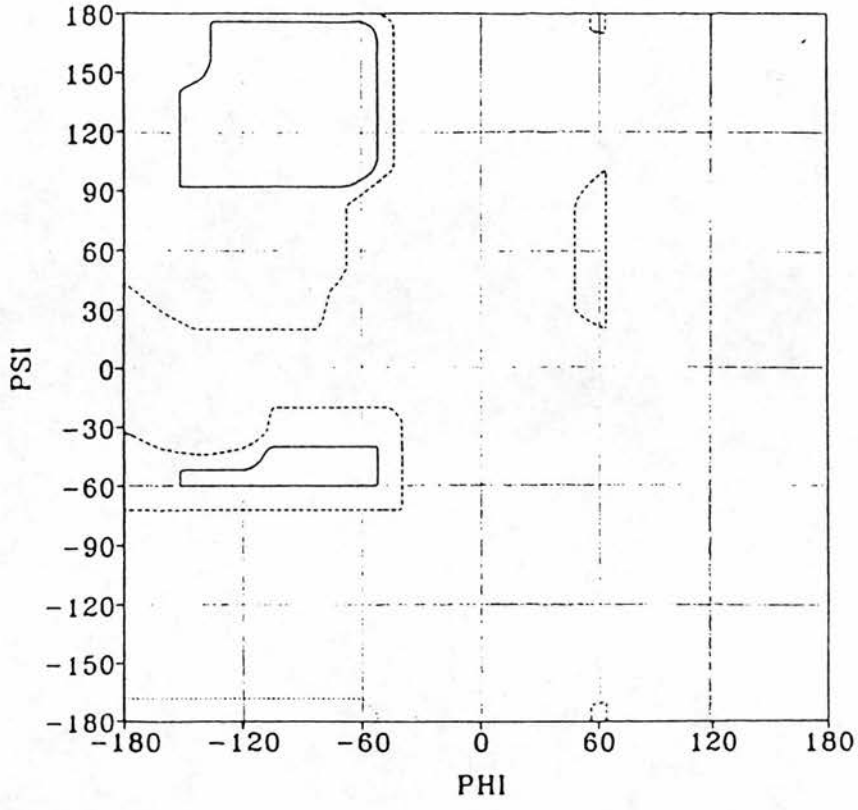


Figure 4–2: Energetically favourable regions of torsional space for protein backbone dihedral angles.

a variety of different ways: size, charge, polar/apolar, acidic, basic, etc. The tertiary structure is dependent on the local interaction of the main chain atoms to form secondary structure elements and long range interaction of the side chain moieties of these secondary structure elements with one another and the solvent. The presence of a beta carbon atom in a side chain constrains the backbone to limited dihedral angles (these are defined in figure 4–2). These limited dihedral angles include those required for the formation of the major secondary structures: α-helix and β-sheet.

Secondary Structure (α-helices and β-strands)

The prominent secondary structures seen in protein structures are right handed α-helices and extended β-strands, which usually associate into β-sheets. The

stability of these structures arises from hydrogen bonding between main chain amino and carbonyl groups. Right-handed α-helices have 3.6 residues per turn and a translation of 5.41 Å per turn. A more tightly coiled helix is occasionally seen in protein structures: the 3_{10} helix. This has approximately three residues per helix turn and has been observed at the end of α-helices, where one turn may have this 3_{10} conformation. A more loosely coiled helix with approximately five residues per turn is theoretically possible, the π-helix, but this has not been observed to date. Beta strands associate to form β-sheets, both parallel and anti-parallel interaction of strands is observed. These sheets have a right-handed twist to the backbone which is favoured energetically. Proline is stereochemically incompatible with either the α-helix or β-sheet conformation, but poly(Pro) is capable of forming its own unique all *cis*- and all *trans*-helices: poly(Pro) I and II.

Tertiary Structure

Secondary structure units often associate to form super-secondary structures. These structures are at a higher level than is secondary structure but do not neccessarily constitute entire structural domains, these super-secondary structures can then associate to form the whole tertiary structure. Four anti-parallel α-helices can associate to form a stable structure when packed at an angle of approximately +20° to each other, e.g. hemerythrin. Beta-strands associate to form β-sheets, two of which may pack orthogonally at 90° to one another (the lipocalycins) or aligned at -30° to one another (the immunoglobulin fold). Alpha helices and β-strands can associate to form $\alpha\beta$ units, several of these can then associate to form a parallel β-barrel with the helices lying external to the barrel, e.g. triose phosphate isomerase. For proteins with more than about 200 residues independent units within the structure are seen. These domains could be a feature of the way proteins have evolved - two smaller proteins becoming one physical polypeptide chain by gene insertion. It is also possible that domains are a necessary feature of the protein folding process, with each domain folding separately.

Quaternary Structure

For some proteins, commonly enzymes, the active biological molecule consists of an association of several copies of the same polypeptide or possibly of different polypeptides. The interfaces between different subunits are generally similar to the interiors of individual molecules. The residues are close packed and both hydrophobic and polar interactions are involved. Interactions between subunits are either isologous or heterologous. In the former, the interacting surfaces are identical, which gives rise to a closed dimeric structure with a twofold axis of symmetry. In heterologous association the interfaces are not identical but are complementary. This kind association can be open-ended giving rise to long assemblies such as F-actin, or the geometry can be such that a closed cyclic structure is produced. It is often the case that full biological activity is only present after quaternary association of subunits.

4.2.2 Interatomic Forces

Protein structure is dependent upon the covalent and non-covalent interactions between atoms. The structural heirarchy outlined above has been elucidated by analysis of high resolution protein crystal structures but has to be understood in terms of the forces acting between atoms.

Covalent Interactions

The covalent bonds within a protein structure are determined by the primary amino acid sequence. The need to maintain bond lengths close to their optimum value places many constraints on the backbone conformation and also the packing of amino acids together. In addition there are preferred, lowest energy conformations for the angles between three atoms and four atoms which are linked covalently (figure 4–3)

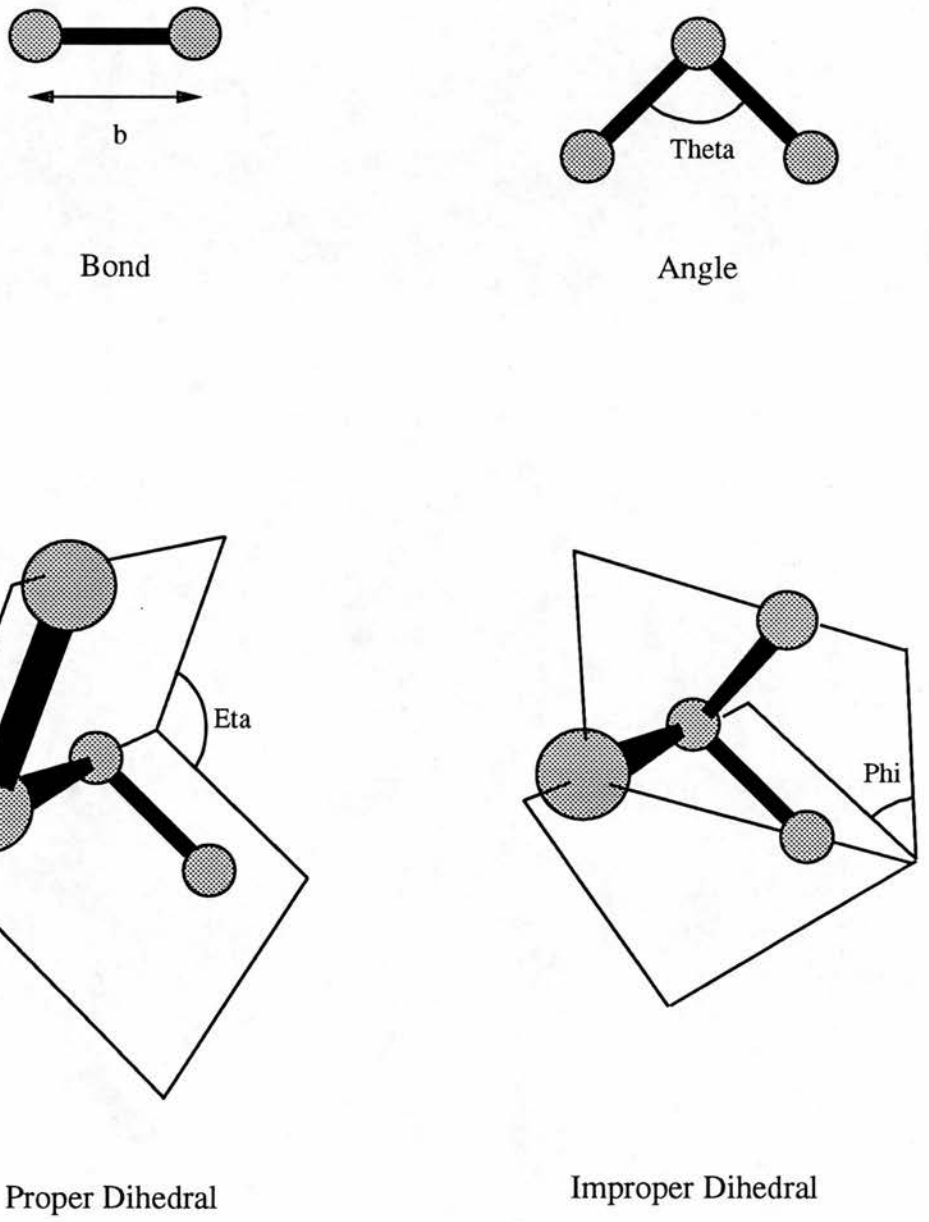


Figure 4–3: Covalent bonds in proteins. Angles and distances shown correspond to those in the force field equations. The improper dihedral is not a physical bond but is defined in order to retain the handedness at certain chiral centres.

Non-covalent Interactions

The forces that stabilise the three dimensional structure of a protein are the through-space interactions between atoms. Alpha-helix and β-sheet stability are dependent on the non-covalent interactions between backbone oxygen, hydrogen and nitrogen atoms.

Dispersion and Electron Repulsion Short range dispersive forces between pairs of atoms are attractive. The movement of electrons around the nucleus creates an oscillating dipole. When two such dipoles are close enough together the oscillation of dipoles becomes coupled, as each dipole polarizes the opposing atom. This creates a weak attractive force between the pair of atoms. When the atoms are part of larger molecules the dispersive force becomes orientation dependent, but is usually considered to be isotropic. The attractive force is counterbalanced by the repulsion of electron shells at small distances. The Lennard-Jones approximation is often used for energy calculations involving proteins, coupled with London's term for dispersive forces;

$$E = \frac{A}{r^{12}} - \frac{B}{r^6} \tag{4.2}$$

where r is the separation between two atoms and A and B are the Lennard-Jones and London constants for that pair of atoms. The magnitude of these parameters, A and B, can be determined from non-bonded distances and contact energies in small molecule crystals. Alternatively they can be found from beam scattering experiments, infrared spectroscopy and heats of sublimation.

Electrostatic Interactions The formation of covalent bonds between atoms can produce a charge asymmetry, one atom being more negatively charged than the other. Therefore, most atoms of a protein carry partial charges as dipoles or higher multipoles. The interaction energy between two partial charges, q_1 and q_2, can be approximated using Coulomb's law

$$E = \frac{q_1 q_2}{4\pi\epsilon_0\epsilon_r r} \tag{4.3}$$

where ϵ is the dilectric constant of the space separating the charges. The values of these partial charges are difficult to calculate and many different methods exist: *ab initio* molecular orbital and quantum mechanical calculations being most common. The agreement between different methods is often poor. The Coulombic interaction described in eqn 4.3 is long range, therefore the electrostatic forces between the many atoms of a protein molecule would appear to be very complex. However, the assumption is made that no free charges exist in proteins, only dipoles or multipoles. The energy of interaction between two dipoles, μ_1 and μ_2, decreases with at least the third power of the inverse distance;

$$E = \frac{\left[\left(\frac{\mu_1.\mu_2}{r^3}\right) - \frac{3(\mu_1.r).(\mu_2.r)}{r^5}\right]}{4\pi\epsilon_0\epsilon_r} \tag{4.4}$$

Therefore, the range of electrostatic interactions in proteins is usually assumed to be short. This is valid to a point but often is only used as an approximation for computational reasons, in fact many proteins function by the movement of electrons from one functional group to another.

Salt Bridges When two charged species interact electrostatically, a salt bridge is formed. The energy for this interaction can be calculated using Coulomb's law as above. When charged groups are exposed to solvent the formation of salt bridges is favoured for entropic reasons. Burying a free charge inside the protein interior is very unfavourable.

Hydrogen Bonds Hydrogen bonds are mainly electrostatic in nature. When a hydrogen atom has a positive partial charge (hydrogen bond donor) and is close to a group with a negative partial charge (hydrogen bond acceptor) hydrogen bonding can occur. The charges attract each other, but the repulsive force due to electron cloud overlap is small because hydrogen only has one electron. The attracting charges can come close together, the short distance gives rise to a high

Coulomb and dispersive energy. The energy of the interaction for hydrogen bonds lies somewhere between the energies of van der Waals contacts and covalent bonds. Hydrogen bonds are linear because they involve a partial positive charge aligned between two partial negative charges; the lowest potential state for the positive charge is when all three charges are aligned. Analysis of protein structures shows deviations of up to 20° in linearity. This deviation is presumably due to a compromise between maximum hydrogen bonding energy and other geometric constraints. Hydrogen bonding is very important in maintaining protein structure.

Entropic Forces (the Hydrophobic Force) At first glance it may seem strange that proteins fold at all. In going from a random disordered structure to a very complex order there is a very high entropic expense. Considering a protein *in vacuo*, folding to a stable ordered state can only occur if the gain in binding energy ΔH is greater than the loss in entropic energy $-T\Delta S$. Obviously as the temperature increases this binding energy must be increased. However, proteins do not fold *in vacuo*, they fold in an aqueous solvent. In a polar medium (eg. water), the solvent entropy greatly stabilises the folded structure of a protein. Considering any generalised apolar molecule in water. The water molecules form an ordered structure around the molecule in order to maintain hydrogen bonds amongst themselves (a negative ΔH but a more negative ΔS). Therefore, the fewer apolar groups that are exposed to the solvent the lower the entropic cost for the order produced in the solvent. In the protein the polypeptide chain is rather more complex, consisting of both polar and apolar groups. But as is predicted from the general case the majority of apolar side chains are removed from interaction with the solvent and become buried in the protein interior. However, the polypeptide backbone has polar amide and carbonyl groups. These can only be buried if the loss of free energy is minimised by the majority of them forming hydrogen bonds. Indeed, in the interiors of protein nearly all the available hydrogen bond donors and acceptors are satisfied (Hubbard and Baker, 1984).

4.3 Calculation of Protein Energetics

Theoretically it is possible to derive the ground state structure of a protein from quantum mechanical calculations. Unfortunately, such calculations are still only possible on structures of less than 100 atoms, due to the massive number of calculations involved (which increases with the cube of the number of atoms). However, the forces governing protein structure are known therefore it is possible to calculate the energetics of static protein structure and also simulate the motion of proteins. These calculations use a Newtonian interpretation of protein structure; atoms are considered as spheres connected by springs. Both of these calculations are based on a potential function which approximates the potential energy of the protein molecule. An in depth background to macromolecular dynamics and minimisation is given by McCammon and Harvey (1987), while an up-to-date review of applications is given by Petsko and Karplus (1991).

4.3.1 The Potential Function

The previous information implies that the potential energy of a molecule has contributions from both covalent and non-covalent interactions. The covalent potential is derived from bond length, bond angle, proper dihedral and improper dihedral angle terms (figure 4–3). Ideal values for these terms for many atom types have been derived from small molecule crystal data and infrared spectroscopy. The covalent potential is calculated from the deviation of terms from these ideal values;

$$E_{bond} = \sum^{bonds} \frac{1}{2} K_b (b - b_0)^2 \tag{4.5}$$

$$E_{angle} = \sum^{angles} \frac{1}{2} K_\theta (\theta - \theta_0)^2 \tag{4.6}$$

$$E_{proper\ dihedral} = \sum^{proper} \frac{1}{2} K\xi (\xi - \xi_0)^2 \tag{4.7}$$

$$E_{improper\ dihedral} = \sum^{improper} K_\phi [1 + \cos(n\phi - \delta)] \tag{4.8}$$

The non-covalent potential consists of the Lennard-Jones repulsive and dispersive energies and the Coulombic energy between atoms. An explicit hydrogen bonding term is sometimes included but is not usually needed as hydrogen bonds are mainly electrostatic in character.

$$E_{van\ der\ Waals} = \sum_{i=1}^{atoms} \sum_{j>i}^{atoms} [(\frac{A}{r^{12}}) - (\frac{B}{r^6})] \tag{4.9}$$

$$E_{coulombic} = \sum_{i=1}^{atoms} \sum_{j>i}^{atoms} \frac{q_i q_j}{(4\pi\epsilon_0\epsilon_r r_{ij})} \tag{4.10}$$

The parameters A and B for van der Waals interaction between pairs of atoms have been calculated from small molecule crystal data and beam scattering experiments. The partial charges q_i and q_j can be calculated as outlined above, but with some inaccuracy. Therefore, the electrostatic term of the force field is the least accurate, but this is not a major problem if it assumed that the electrostatic interactions are short range. Given this description of the potential energy for a protein molecule, and the bonded and non-bonded parameters, it is possible to calculate the potential energy for any protein conformation. This would be interesting in itself, to compare different possible conformations of the same molecule. However, the problem can be further expanded. By considering the first derivative of the potential V with respect to the atomic coordinates $\mathbf{x}$;

$$-V'(\mathbf{x}) = \text{Force} = \text{ mass.acceleration} \tag{4.11}$$

a minimum of V occurs when:

$$V'(\mathbf{x}) = 0 \tag{4.12}$$

The first derivative of the potential with respect to position can therefore be used to calculate a minimum of the potential energy and also the acceleration on individual atoms.

4.3.2 Energy Minimisation

Computationally, energy minimisation is a problem of nonlinear optimization. In the protein case there is a set of independent variables $\mathbf{x}$, the atomic coordinates, and an objective function $V(\mathbf{x})$, the potential, derived from $\mathbf{x}$. The problem is to find the set of values for $\mathbf{x}$ which produces a minimum of $V(\mathbf{x})$. There are different algorithms for searching for this minimum. They can be categorised in terms of the highest order derivative of the potential function which is used. In general a function $f(x)$ can be expanded as a Taylor series about the point x_0,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2 f''(x_0)}{2} + \ldots \tag{4.13}$$

Grid Search (Zero Order Method) The minimum potential is found by scanning the possible values for the independent variables $\mathbf{x}$ in a systematic way. This can be first done on a coarse sampling grid, low potential points being further sampled with a finer grid. This method can be successfully used only when the number of independent variables is small and the potential surface is relatively simple. As a consequence this method is very rarely considered for protein molecules.

First Order Methods The first derivative of the potential energy describes the local gradient of the potential energy surface. The first derivative is in fact the negative of the force;

$$F = -V'(\mathbf{x}) \tag{4.14}$$

Because the potential function is an explicitly differentiable function of the atomic coordinates, the force on each atom can be readily calculated. The atoms can be moved to minimise the force acting on each atom. Two methods are commonly used: steepest descent and conjugate gradient, both of which are iterative descent techniques.

Steepest Descent The protein conformation just prior to the kth minimisation step is specified by the $3N$ dimensional vector $\mathbf{x}_{k-1}$, where N is the number of atoms involved. A descents direction is chosen, represented by another $3N$ dimensional vector of unit length $\mathbf{s}_k$. A scalar descent step size, λ_k, is determined. The descent step is taken by;

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \lambda_k \mathbf{s}_k \tag{4.15}$$

The descent direction is parallel to the net force;

$$\mathbf{s}_k = \frac{F}{|F|} \tag{4.16}$$

The step size λ_k is usually determined by a simple criterion. The initial step size is chosen and a step taken. If the energy is reduced by this step the step size is increased by some multiplying factor (typically 1.2) for the next iteration. This increase in step size occurs as long as the iteration reduces the potential energy. When the step produces an increase in energy the step size is reduced typically by a factor of 0.5. This iterative change in step size serves to make successive descents steps of a size appropriate to the shape of the potential energy surface. The steepest descent algorithm performs well far from a minimum, reaching a local minimum rapidly. Unfortunately it is a nonconvergent method and is inefficient for problems with many local minima, such as proteins. It quickly eliminates the worst steric clashes and brings bond lengths and angles to approximately optimum values, but it cannot produce the collective motions, due to weak forces on many atoms, that are necessary to generate an optimum overall structure. A more efficient technique is the conjugate gradient algorithm.

Conjugate Gradient This method combines information about the present gradient with gradients from previous steps to determine the descent direction. The initial search direction is taken along the negative gradient;

$$\mathbf{s} = -\mathbf{g}_1 \tag{4.17}$$

At the next step the search direction is determined by;

$$\mathbf{s}_k = -\mathbf{g}_k + b_k \mathbf{s}_{k-1} \tag{4.18}$$

where the parameter b_k is a weighting factor equal to the ratio of the squares of magnitude of the current and previous gradient;

$$b_k = \frac{|\mathbf{g}_k|^2}{|\mathbf{g}_{k-1}|^2} \tag{4.19}$$

It can be shown that for an n dimensional surface this method will pass through the minimum on the nth step, provided the minimum along each successive search direction is found. Even when the step size is not optimum, the conjugate gradient method produces search directions that are superior to the steepest descent method. Despite the conjugate gradient method requiring two evaluations of the gradients per minimisation step it is more efficient than the steepest descent method which requires only one. It should be noted that convergence in n steps is based on the assumption that the potential energy surface is quadratic; this is, in general, not the case. The accumulated errors in the search direction are therefore removed every m steps by setting b_m equal to zero.

Newton-Raphson method (second order) This method assumes that the potential depends approximately quadratically on the independent variables, for a quadratic function $f(x)$:

$$f(x) = a + bx + cx^2 \tag{4.20}$$

Therefore,

$$f'(x) = b + 2cx \tag{4.21}$$

$$f''(x) = 2c \tag{4.22}$$

At the minimum, $f'(x) = 0$ therefore

$$x_{min} = -b/2c \tag{4.23}$$

substituting

$$x_{min} = x - f'(x)/f''(x) \tag{4.24}$$

For quadratic energy surfaces no iteration is required, the minimum is calculated directly from the given configuration x. Unfortunately, the highly nonquadratic nature of the protein potential surface and the many local minima make the use of the Newton-Raphson method unsuitable. In addition, the need to invert the second derivative matrix is computationally expensive, as for N atoms it is $3N \times 3N$ in size. It is noted that a modified technique, adopted basis set Newton-Raphson minimisation, has been used successfully in the CHARMM package (Brooks *et al.*, 1983).

4.3.3 Molecular Dynamics Simulations

If it is assumed that each atom is a point mass whose motion is determined by the forces exerted upon it by all the other atoms in the system, this motion can be described by the equations of motion of classical mechanics. These equations must be solved numerically for three or more independent particles. We have seen already that the force on an atom i is given by the negative gradient of the potential function with respect to the position of particle (eqn. 4.11). Considering the motion of a generalised single atom along a particular coordinate, $\mathbf{x}$. If the position at time t is known, the position after a short time step δt is given by the Taylor expansion;

$$\mathbf{x}_{t+\delta t} = \mathbf{x}_t + \mathbf{x}'_t \delta t + \frac{\mathbf{x}''_t \delta t^2}{2} + \ldots \tag{4.25}$$

The numerical solution of the equations of motion requires the position, velocity and acceleration at time t to be known. Suitable approximations to account for contributions from higher derivatives are made so that $\mathbf{x}_{t+\delta t}$ can be calculated with reasonable precision. The instantaneous acceleration $\mathbf{x}''$ on particle i is given by Newton's second law;

$$\mathbf{x}_i'' = \frac{F_i}{m_i} \tag{4.26}$$

The Verlet Method

If $\mathbf{v}$ is the average velocity across the time interval between t and $t + \delta t$ then;

$$\mathbf{x}_{t+\delta t} = \mathbf{x}_t + \mathbf{v}\delta t \tag{4.27}$$

Assuming that $\mathbf{v}$ is very nearly equal to the velocity at the midpoint of the time interval;

$$\mathbf{v} = \mathbf{x}'_{t+\frac{\delta t}{2}} \tag{4.28}$$

This can be calculated if $\mathbf{a}$, the average acceleration during the interval $t - \frac{\delta t}{2}$ to $t + \frac{\delta t}{2}$ is known;

$$\mathbf{x}'_{t+\frac{\delta t}{2}} = \mathbf{x}'_{t-\frac{\delta t}{2}} + \mathbf{a}\delta t \tag{4.29}$$

Assuming that $\mathbf{a}$ is very nearly equal to the instantaneous acceleration at the midpoint of this interval;

$$\mathbf{a} = \mathbf{x}_t'' \tag{4.30}$$

Therefore

$$\mathbf{x}'_{t+\frac{\delta t}{2}} = \mathbf{x}'_{t-\frac{\delta t}{2}} + \mathbf{x}''_t \delta t \tag{4.31}$$

Substituting,

$$\mathbf{x}_{t+\delta t} = \mathbf{x}_t + \left(\mathbf{x}'_{t-\frac{\delta t}{2}} + \mathbf{x}''_t \delta t\right) \delta t \tag{4.32}$$

The acceleration is calculated from the force F at time t. This algorithm is often called the leapfrog method because the velocity is calculated at odd half integral multiples of δt, while the position is calculated at integral multiples of δt.

The Constrained Verlet (SHAKE) Method

A major problem with molecular dynamics simulations on large systems is the computational expense. Therefore, any method that can be used to decrease the computation required is desirable. An algorithm based on novel parallel architecture is discussed in chapter 5. Other computational approaches have included increasing the time step δt. The maximum time step is determined by the requirement that δt be small in comparison to the period of the highest frequency motion in the system being simulated. In proteins the highest frequency motions are the bond stretching vibrations involving hydrogen atoms. Removing these vibrations by constraining the bond lengths to a fixed length enables a longer time step to be used. The kth constraint on the distance between two atoms i and j can be expressed;

$$\mathbf{r}_{ij}^2 - d_{ij}^2 = 0 \tag{4.33}$$

where $\mathbf{r}_{ij}$ is the vector between the cartesian coordinates of i and j;

$$\mathbf{r}_{ij} = \mathbf{r}_j - \mathbf{r}_i \tag{4.34}$$

and d_{ij}^2 is the correction that must be applied to $\mathbf{r}_{ij}^2$ for the constraint to be satisfied. In a numerical simulation method all constraints cannot be satisfied

exactly, therefore the constraint is said to be satisfied when the relative deviation is below a set limit ϵ;

$$s_k = \frac{(\mathbf{r}_{ij}^2 - d_k^2)}{d_k^2} < \epsilon \tag{4.35}$$

Given a set of coordinates $\mathbf{x}$, which satisfy the constraints, a Verlet integration step can be taken to yield a set of coordinates $\mathbf{x}^*$. The problem is to determine the adjustments which need to be applied to $\mathbf{x}^*$ to satisfy all the constraints;

$$\mathbf{x}^{**} = \mathbf{x}^* + \delta\mathbf{x} \tag{4.36}$$

For each pair of constrained atoms it can be shown (McCammon and Harvey, 1987) that the adjustments for the kth restraint are;

$$\delta^k\mathbf{r}_i = -\frac{g_{ij}\mathbf{r}_{ij}}{m_i} \tag{4.37}$$

$$\delta^k\mathbf{r}_j = \frac{g_{ij}\mathbf{r}_{ij}}{m_j} \tag{4.38}$$

The problem is how to solve for g_{ij}. Fortunately, the problem can be reformulated as an equation that can be solved iteratively, provided that the adjustments applied are very much smaller than the constrained bond lengths (van Gunsteren and Berendsen, 1977). Adjustments are successively applied until all constraints satisfy the specified tolerance limit. The application of the SHAKE algorithm enables time steps of up to 2 femtoseconds to be used, rather than 0.5 femtoseconds for unconstrained integration steps. Although SHAKE is an iterative process the increase in computational efficiency can be up to 3 fold.

4.3.4 Scope of Minimisation and Dynamics Simulations

Due to the complexity of a protein potential energy surface, energy minimisation cannot find a global minimum. Therefore the determination of a protein structure *ab initio* cannot be achieved by minimisation techniques. If the motion

of an extended random conformation of a protein could be simulated for a long enough time scale the folded state might be reached. Unfortunately, proteins fold in the time scale of seconds, whereas the longest simulations manageable are in the order of nanoseconds. The limited time scale for simulations is due to the short integration time step required ($\sim$ 1 fs) and the large computation needed to calculate the first derivatives of the potential. Modern super computers can only realistically be used calculate in the order of a million integration steps. Therefore, neither minimisation or molecular dynamics can calculate the folded state for an unknown structure. Both techniques are useful for investigating the properties of structures which have already been determined by crystallography or NMR. In order to approach an unknown structure, by computational methods, the homology modelling techniques outlined earlier must first be applied to arrive at an approximate structure. The energy minimisation and molecular dynamics techniques described can be most usefully employed in optimizing this initial structure.

4.4 Molecular Modelling of a2u

The previous sections in this chapter have suggested that is possible to model a protein's structure from other information, provided this information includes structures of proteins with related sequence and function. The way in which this information can be best used to model a protein's structure is not rigidly defined. This section presents the modelling of a2u using four different strategies, the results of the different methods being compared amongst each other. The validation of model structures is considered with respect to these four models.

4.4.1 Sequences and Structures

Sequences were obtained from the NBRF database using the UWGCG package (Devereux *et al.*, 1984) and checked against recent literature reports. The sequences for RBP, INSC and BLG were pir:vahu (Colantuoni *et al.*, 1983),

	RBP	BLG	INSEC
a2u	46.8 20.2	48.1 23.1	46.0 18.7
RBP	-	41.5 20.7	43.6 20.3
BLG	-	-	43.4 25.6

Table 4–1: Sequence similarity (upper) and identity (lower) for pairwise alignments of lipocalycin crystal structures.

pir:cuwoi (Riley *et al.*, 1984), and pir:lgbo (Braunitzer *et al.*, 1972) respectively. The sequences for a2u and MUP were pir:uart (Dolan *et al.*, 1982) and pir:uams (Clark *et al.*, 1984) respectively. X-ray crystallographic coordinates were obtained from T.A.Jones (RBP) (Newcomer *et al.*, 1984), H.M.Holden (INSEC) (Holden *et al.*, 1987), and L.Sawyer (BLG) (Papiz *et al.*, 1986). The structure of RBP was a preliminary model refined to an R-value of 28% at 2.8 Å resolution. The structure of INSEC was that reported in the literature, refined to an R-value of 13.7% at 2.6 Å resolution. The structure of BLG was the result of GROMOS-MDXREF refinement at 3.0 Å, with a final R-value of 32%. Further atomic coordinates were obtained from the Brookhaven Protein Databank (PDB) (release 57; Bernstein *et al.*, 1977).

4.4.2 Sequence Alignment

Pairwise sequence alignments were carried out using the GAP program from within UWGCG. This program uses the Needleman and Wunsch algorithm, described earlier, to align two sequences. Multiple sequence alignments were carried out using the program CLUSTAL. This program uses cluster analysis of the results of several pairwise alignments to multiply align several sequences. Pairwise alignments between all sequences were carried out using the GAP program. This gave an indication of the similarity of the sequences, approximately 20% identity and 45% similarity (table 4–1).

4.4.3 Structural Alignments

Earlier it was suggested that the level of identity between protein sequences could determine the modelling strategy. If a sequence shows high sequence identity (¿ 70%) to a known structure it is possible to change the side chains of this structure to those of the other sequence and arrive at a model close to the crystal structure, for example trypsin and chymotrypsin. However, if a sequence shows limited identity (30% or lower) to several structures it is desirable to include the information from all these structures in the modelling procedure. One obvious approach is to align the structures in space so that conserved regions in the sequence are coincident. This alignment can show which regions of the sequence are conserved structurally, giving a core structure around which a model can be constructed. The most common method for optimising the superposition of one coordinate set to another is least squares fitting (Lesk, 1991). Given two sets of coordinates, each with the same number of atoms, a residual function can be defined that expresses the closeness of fit of n equivalent atoms:

$$R = \sum_{1}^{n} w_i(\mathbf{x}_i - \mathbf{x}_i^*)^2 \tag{4.39}$$

where $\mathbf{x}$ and $\mathbf{x}^*$ are the two coordinate sets, and w is a weight for each coordinate pair. The position of one coordinate set relative to the other can be defined by a set of m parameters, $p_j = 1, m$. The least squares procedure is a method for adjusting these parameters in such a way as to give a best fit between the two coordinate sets. In the case of rigid-body superposition of two structures there are only six parameters: three rotational and three translational. In this context best is defined as the set of parameter values (the orientation of the second molecule relative to the first) that minimises the residual function. The minimum can be defined:

$$\frac{dR}{dp_j} = \sum_{1}^{n} -2w_i(\mathbf{x}_i - \mathbf{x}_i^*)\frac{d\mathbf{x}^*}{dp_j} = 0 \tag{4.40}$$

for all m values of j. The minimum can be determined directly if the the expressions defined above are linear in terms of the parameters 1 through to m. Alternatively the minimum can be found using iterative search or gradient techniques (see previous section). This technique for the superposition of two coordinate sets has been implemented in several graphics programs; HYDRA (Hubbard, 1985), SYBYL (Tripos Associates, 1991), and O (Jones, 1991). Both HYDRA and SYBYL require the user to state explicitly which atoms are to be fitted, usually main chain or alpha carbons only. This obviously introduces an element of subjectivity into the procedure as the user must decide which regions of the structures are similar enough to allow a good fit to be produced. The program O extends the procedure to allow optimisation of the first explicit matching of atom pairs. The program uses a search for structure fragments that lie within some cutoff distance whose sequences also align well. The algorithm is iterative so that larger and longer fragments are included as the fit between the structures inproves (O manual).

4.4.4 Loop Searches

The most variation seen in structures with similar sequences is in the loop regions. That is to say the core secondary structure elements are well conserved and constant but the regions joining these elements are variable. This is to be expected as the loop regions joining secondary structure elements are often exposed to solvent, therefore variation in their size and amino acid content, provided they remain hydrophilic in nature, are possible. It is much more difficult to make large changes to the number and type of residues in the core of a protein and maintain the same overall structure (Chothia and Lesk, 1986). During the process of homology modelling it is quite often the case that the core protein structure can be modelled with some degree of accuracy but the loop regions are ill-defined. The sequence alignment of a group of related proteins often shows that insertions or deletions, where greater or fewer amino acids must be accommodated respectively, occur in those regions between secondary structure elements which are exposed to the solvent. Therefore, the problem is to

determine the structure of these loop regions in some way. A procedure has been developed which makes use of existing structural information. It is assumed that the coordinates of the residues that start and finish a loop and the number of amino acids in the loop itself determine the structure of the loop to some degree. It is possible to search a database of the known structures for amino acid fragments which have similar start/end points and amino acid lengths (figure 4–4). If this process is carried out it becomes clear that there are many different loop structures for a given set of parameters. This introduces another element of subjectivity in the modelling process, with the user deciding which loop best fits the structure. Improved algorithms for automatically determining the best loop from a group of possible candidates are being developed, but are not well established (Summers and Karplus, 1990).

4.4.5 Energy minimisation

The modelling of a protein structure from other structures will usually produce a model with unfavourable non-bonded interactions. As suggested earlier energy minimisation and molecular dynamics simulations can be used to remove these bad contacts and optimise the model. There are many minimisation/simulation programs available; GROMOS87, X-PLOR, AMBER, CHARMM, being those most widely used. These packages often use the same underlying methods to minimise and simulate proteins and other macromolecules. There are sometimes differences between the forcefield parameters used by different programs especially when nonstandard chemical groups are used. However, many of the parameters for the standards amino acids are similar for each program, being derived from one common source (Weiner *et al.*, 1984; Weiner *et al.*, 1986). For the modelling of a2u the GROMOS87 package (van Gunsteren, 1987) and a program which implemented the GROMOS87 forcefield on a parallel computer (chapter 5) were used. Energy minimisation was carried out using the PROEML program from GROMOS87, and a parallel implementation of the PROEML steepest descents minimiser. The same protocol was followed for the minimisation of all the models (figure 4–5).

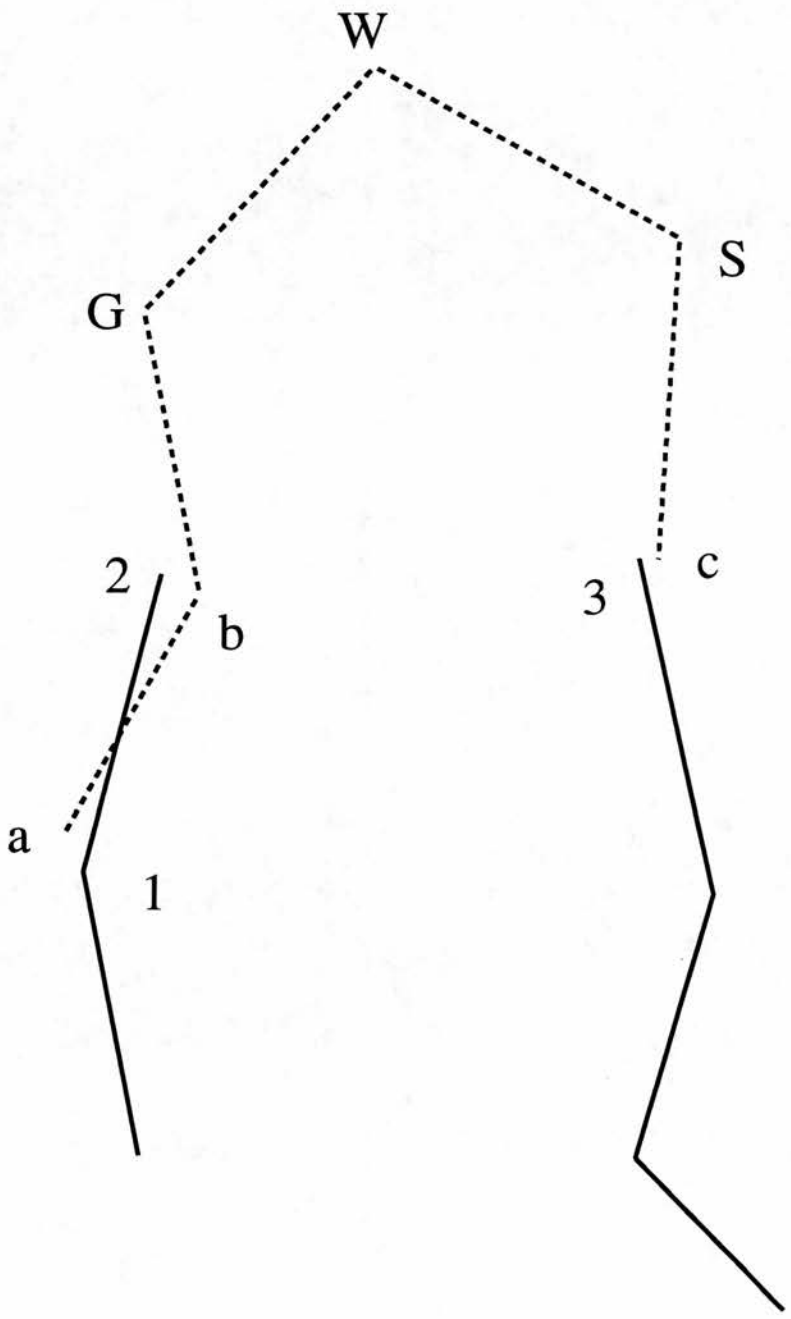


Figure 4–4: Method of loop searching. Points a, b, c and 1, 2, 3 can be superimposed with a low rms deviation.

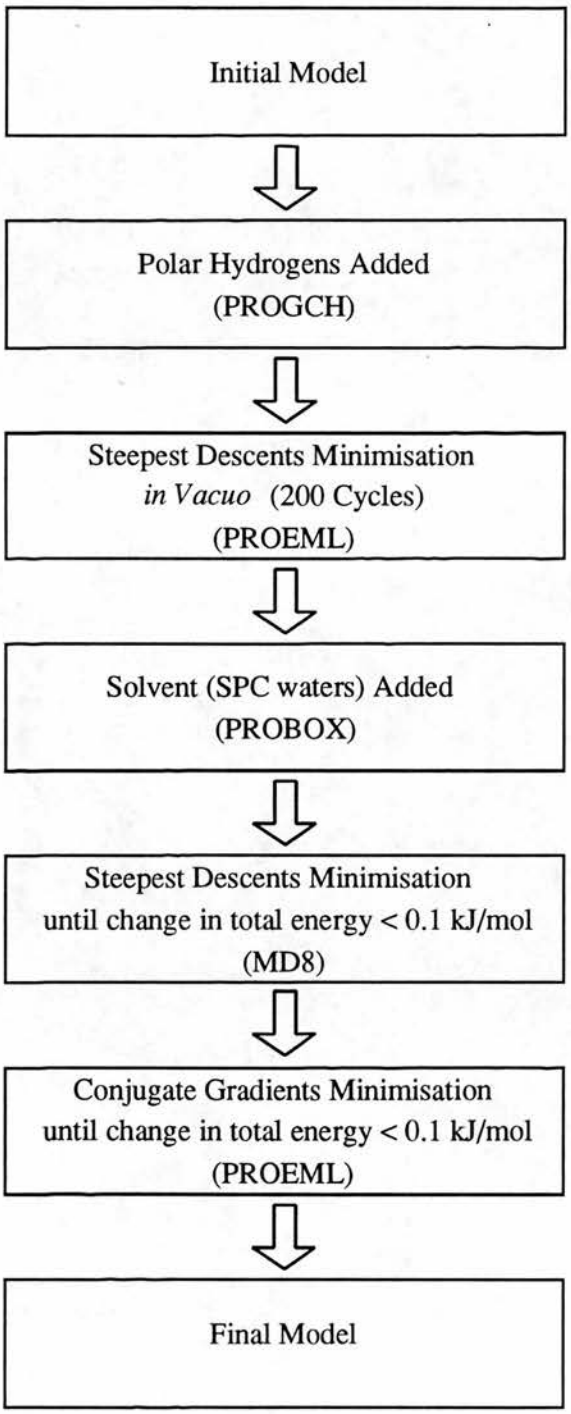


Figure 4–5: Energy minimisation protocol used for modelling.

```
a2u     EEASSTRGNLDVAKLNGDWFSIVVASNKREKIE.ENGSMRVFMQHIDVL.
          .. |. .||:.|:.|.|:|:.:|... . :: :.:.:||:::.:..
BLG     LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTP

a2u     ENSLGFKFRIKENGECRELYLVAYKTPEDGEYFVEYDGGNTFTILKTDYD
        |..|:: :. .||||| :  ::| ||. .: : ::  .:|.. :|.|||.
BLG     EGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYK

a2u     RYVMFHLINFKNGETFQLMVLYGRTKDLSSDIKEKFAKLCEAHGITRDNI
        :|::| : |  :.|       ..||.::..:  |||.|  . .::...
BLG     KYLLFCMENSAEPEQSLACQCLVRTPEVDDEALEKFDK..ALKALPMHIR

a2u     IDLTKTDRCLQARG
        :.:..|:   |.:
BLG     LSFNPTQLEEQCHI
```

Figure 4–6: Pairwise sequence alignment for a2u and BLG. Alignment was carried out with the UWGCG program GAP using default parameters.

4.4.6 Method 1

From table 4–1 It can be seen that BLG has the highest sequence indentity and similarity score to the sequence of a2u. A pairwise alignment of these two sequences shows many identical and conserved residues, only two deletions (each one amino acid) and one insertion (two amino acids) are needed for the alignment (figure 4–6). It was relatively straight forward to mutate the BLG sequence to that of a2u. Insertions and deletions were in loop regions in the BLG structure. This was all carried out using the REPLACE option in FRODO. It was then possible to manipulate the protein backbone manually using the .TOR option in FRODO. The .TOR option allowed main chain torsional angles to be rotated using interactive dials. These rotations were performed until the gap between mainchain nitrogen and carbon atoms was of a bonding distance. It was therefore possible to close gaps where residues had been deleted and remove steric clashes by direct manipulation of the backbone.

2° Structure Element	Residues		
	BLG	RBP	INSEC
$\beta 1$	17-27	22-31	25-32
$\beta 2$	38-43	37-48	42-51
$\beta 3$	47-53	52-62	54-63
$\beta 4$	69-75	67-78	66-76
$\beta 5$	81-84	84-91	84-94
$\beta 6$	88-97	99-110	97-109
$\beta 7$	102-109	113-122	112-122
$\beta 8$	115-124	131-139	127-137
$\alpha 1$	130-140	145-160	143-157
$\beta 9$	144-149	166-169	166-169

Table 4–2: Residues forming the core secondary structure elements for BLG, RBP and INSEC.

4.4.7 Method 2

The structures of RBP, BLG and INSEC were superimposed using the least squares fitting subprogram in HYDRA. It was seen that the common structural elements do not neccessarily coincide with the common sequence elements. The secondary structure elements were identified visually for each structure. As described in chapter 2 this gives a common core structure of eight β-strands followed by one α-helix and one further β-strand (table 4–2 and figure 4–7). These structural elements were combined with sequence information (figure 4–8 to model a core structure for a2u. Alignment of the a2u sequence with the other sequences within secondary structure elements indicated from which structure that element should be taken. The loop regions between secondary structure elements were filled using the sequence alignments or by searching the PDB database with the relevant a2u sequences (table 4–3). Any gaps were closed manually as before using FRODO.

4.4.8 Method 3

The a2u sequence was modelled to the RBP structure in a manner similar to method 1. The biological function of RBP and a2u show some similarities, particularly the renal catabolism of both proteins. This similarity in biological

Figure 4–7: Core secondary structure elements of RBP (red), BLG (green), and INSEC (blue), superimposed using HYDRA.

Structural Element	Source		
	PDB file	Residues	Sequence
N-terminus	RBP	1-8	ERDCRVSS
β1	BLG	9-28	GLDIQKVAGTWYSLAMAASD
loop	RBP	34-40	GLFLQDN
β2	BLG	37-44	APLRVYVE
loop	BLG	45-49	ELKPT
β3	BLG	51-54	EGDL
loop	RBP	62-64	RLL
	6ADH	119-125	PRGTMQD
β4	BLG	62-76	ENGECAQKKIIAEKT
loop	RBP	80-82	TED
β5	RBP	84-90	AKFKMKY
β6/β7	BLG	87-118	LNENKVLVLDTDYKKYLLFCMENSAEPEQSLV
β8	BLG	122-128	LVRTPEV
loop	INSEC	139-141	KVL
α1	BLG	130-138	DEALEKFDK
loop	1ABP	289-297	ITRDNPKEEL
β9	RBP	159-164	LCLARQ
C-terminus	BLG	158-162	EQCHI
	INSEC	183-189	LTGPDRH

Table 4–3: Source of coordinates used to model a2u by method 2.

```
BLG       ......LIVTQTMKGLDIQKVA..|GTWYSLAM|AASDISLLDAQSAPL.
RBP       ...ERDCRVSSFRVKENFDKARFS|GTWYAMAK|..KDPEGLFLQ.....
INSEC     GDIFYPGYCPDVKPVNDFDLSAFA|GAWHEIAK|...LPLENENQG....
a2u       EEASSTSGNLDVAKLN........|GDWFSIVV|.ASNKREKIEENGSM.
                                     b1

BLG      |RVYV|EEL........|KPTPEGD|LEILLQKWENGECAQ|KKIIAEK|.
RBP      |DNIV|AEFSVDETGQM|SATAKGR|.VRLLNNWD......|VCADMVG|T
INSEC    |KCTI|AEYKYDGK...|KASVYNS|..FVSNG........|VKEYMEG|.
a2u      |RVFM|QH.........|IDVLENS|LGFKFRIKENGECRE|LYLVAYK|T
           b2                 b3                        b4

BLG      ...TKIPA........|VFKI|....DAL.....|NENKVLVLDT|DYKK
RBP      FTDTEDPA........|KFKM|KYWGVASFLQK.|GNDDHWIVDT|DYDT
INSEC    ..DLEIAPDAKYTKQG|KYVM|....TFKFGQRV|VNLVPWVLAT|DYKN
a2u      PEDG............|EYFV|EYD.........|GGNTFTILKT|DYDR
                          b5                    b6

BLG      ..|YLLFCME|NSAEPE....|QSLACQCLV|.RTPEVD.|DEALEKFDK
RBP      YA|VQYSCRL|..LNLDGTCA|DSYSFVFSR|..DPNGLP|PEAQKIVRQ
INSEC    YA|INYNCDY|...HPDKKAH|SIHAWILSK|SKVLEGNT|KEVVDNVLK
a2u      ..|YVMFHLI|NFKNG.....|ETFQLMVLY|GRTKDLS.|SDIKEKFAK
               b7                  b8                  a1

BLG      AL|KALPMHI..|RL|SFNPTQLEEQCHI............
RBP      RQ|EELCLARQY|RL|....IVHNGYCDGRSERNLL.....
INSEC    TF|SHLIDASKF|IS|....NFSEAACQYSTTYSLTGPDRH
a2u      ..|..LCEAHG.|IT|RDNIIDLTKTDRCLQARG.......
                     b9
```

Figure 4–8: Combined secondary structure and sequence alignment for RBP, BLG, INSEC and a2u.

```
a2u     .....EEASSTRGNLDVAKLNGDWFSIVVASNKREKIEENGSMRVFMQHI
              .. ..:::|:| |::.|.|:.:.  . .   :::|    . ::..
RBP     ERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDET

a2u     DVLENSLGFKFRIKENGECRELYLVAYKTPED.GEYFVEYDGGNTFT...
        : :...   :.|: :| :. .  :..:...|| :.: :.| |...|
RBP     GQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKG

a2u     .....ILKTDYDRYVM...FHLINFKNGETFQLMVLYGR.TKDLSSD...
             |:.|||| |.:    :|:|:... .  . .:::| ..:|..:
RBP     NDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQK

a2u     .IKEKFAKLCEAHG...ITRDNIIDLTKTDRCLQARG
         :::: ..|| |:.   |.::...| .:.:| |
RBP     IVRQRQEELCLARQYRLIVHNGYCD.GRSERNLL...
```

Figure 4–9: Pairwise sequence alignment of a2u and RBP. Alignment carried out using the UWGCG program GAP with default parameters.

profile and a similar sequence similarity score to that for a2u and BLG suggested that using RBP as a modelling basis may be valid. In addition the modelling of a2u using BLG had raised doubts as to the correctness of the structure of BLG, this will be discussed later. The pairwise alignment of a2u and RBP shows large deletions in the RBP sequence which is some 20 residues longer (figure 4–9). The regions 84-106 and 116-135 in the RBP structure were removed and replaced by structures with close sequence similarity to a2u. Loops were obtained by searching the PDB database with the relevant a2u sequences. These loops were:

- RBP 84-106 replaced by residues 60-74 (chain A) from PDB file 1REI.dat
- RBP 116-135 replaced by residues 59-75 (chain L) from PDB file 3PCY.dat

The last 12 residues of a2u were still undefined as the RBP coordinates only extended to residue 175. These final 12 residues were added in a β-sheet/random conformation as a loop search gave no suitable candidates. The RBP sequence was mutated to that of a2u using the REPLACE command in FRODO. Any gaps were closed manually using FRODO.

4.4.9 Method 4

The sequences of RBP, BLG, INSEC and a2u were aligned using CLUSTAL (figure 4–10). The three structures were superimposed using the BIO FIT command in SYBYL which uses a least squares algorithm to find the best fit for explicitly defined residues. The residues used in the fitting procedure were selected by eye after visual inspection of the three structures, they were: Ser132-Asp140 (RBP), Ile130-Ser138 (INSEC), and Leu117-Thr125 (BLG). A core alpha carbon backbone structure was created from these superimposed structures. A program was written to read in the superimposed structures, the first structure in the list was the template structure. The closest α-carbon atom from all other structures to each α-carbon in the template structure was determined. If at least one atom from another structure lay within a specified cutoff distance of the template structure the template atom and its matches were selected. The centre of geometry of these atoms was calculated for each template α-carbon with at least one match

$$<\mathbf{x}>=\frac{[\sum_{i=1}^{m}\mathbf{x}_i]}{m} \tag{4.41}$$

where m is the number of matches for that template α-carbon (table 4–4). These averaged α-carbon coordinates were written out in Brookhaven format and displayed. This procedure was carried out in order to determine a core structural region for the three x-ray structures (figure 4–11). The core structure produced retained the amino acid sequence of the template structure allowing the residues to be changed to that of a2u on the basis of the multiple sequence alignment. Main chain backbone atoms were rebuilt from these alpha carbon atoms using the CONSTRUCT BACKBONE option in SYBYL. This command uses an algorithm which searches a compressed database of structures for fragments whose α-carbons match closely those in the user model. These fragments are then combined to produce a complete mainchain trace for the model (Claessens *et al.*, 1989). Loop regions were filled using the LOOP command in SYBYL. This searches a compressed database of protein structures for loops with

```
>rbp    ----ERDCRVSSFRVKENFDKARFSGTWYAMAKKDDPEGLFLQDNI--VAEFSVDETGQMS
>blg    ---------LIVTQTMKGLDIQKVAGTWYSLAMAAASDISLLDAQSAPLRVY-VEELKPTP
>insec -GDIFYPGYCPDVKPVNDFDLSAFAGAWHEIAKLPLENEN-QGKCT--IAEY-KYDGKKAS
>a2u    ---------EEASSTRGNLDVAKLNGDWFSIVVASNKREKIEENGS--MRVF-MQHID-VL
                           *     * *

>rbp    ATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYA
>blg    EGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKV--------LVLDTDYKKYL
>insec VYNSFVSNGVKEYMEGDLEIAPDAKYTKQGKYVMTFKFGQRVVNLV---PWVLATDYKNYA
>a2u    ENSLGFKFRIKENGECRELYLVAYKTPEDGEYFVEYDGGNTFT--------ILKTDYDRYV
                                                          ***  *

>rbp    VQYSCRLLNLDGTCADSYSFVF-SRDPNGLPPEAQKIVRQRQE--ELCLARQYRLIVHNGY
>blg    LFCMENSAEPEQ---SLACQCL-VRTPEVDDEALEKFDKALKA------LPMHIRLSFNPT
>insec INYNCDYHP-DKKAHSIHAWIL-SKSKVLEGNTKEVVDNVLKTFSHLIDASKFISNDFSEA
>a2u    MFHLINF-KNGE---TFQLMVLYGRTKDLSSDIKEKFAK-LC---EAHGITRDNIIDLTKT


>rbp    -----CDGRSERNLL-
>blg    QLEEQCHI--------
>insec ACQYSTTYSLTGPDRH
>a2u    ---DRCLQARG-----
```

Figure 4–10: Multiple sequence alignment of RBP, BLG, INSEC, and a2u. Alignment carried out with CLUSTAL using default parameters.

matching start and end points. Those regions not part of the calculated core or well defined loops were resolved by superposition of the known structures onto the model. On the basis of sequence alignment and three dimensional position these regions could be directly transfered from one known structure to the model. In this way the whole model was created (table 4–5). Any bad local geometry was regularised using the ANNEAL option in SYBYL. This is a process of local energy minimisation which regularises covalent geometry considering only local non-bonded interactions.

4.4.10 Energy Minimisation of the Models

All of the models were energy minimised using the protocol outlined in figure 4–5. Steepest descents minimisation *in vacuo* was applied first to remove any major unfavourable energy terms, such as bad non-bonded contacts and irregular bond lengths and angles. Steepest descents was used because of its

RBP	BLG	INSEC
ARG 10	GLN 5	- -
VAL 11	THR 6	- -
LYS 12	MET 7	- -
PHE 15	LEU 10	- -
ASP 16	ASP 11	ASP 19
LYS 17	- -	LEU 20
ALA 18	GLN 13	SER 21
ARG 19	LYS 14	ALA 22
PHE 20	VAL 15	PHE 23
SER 21	ALA 16	ALA 24
GLY 22	GLY 17	GLY 25
THR 23	THR 18	ALA 26
TRP 24	TRP 19	TRP 27
TYR 25	TYR 20	HIS 28
ALA 26	SER 21	GLU 29
MET 27	LEU 22	ILE 30
ALA 28	ALA 23	ALA 31
LYS 29	MET 24	LYS 32
LYS 30	ALA 25	LEU 33
VAL 42	- -	ILE 45
ALA 43	LEU 39	ALA 46
GLU 44	ARG 40	GLU 47
PHE 45	VAL 41	TYR 48
SER 46	TYR 42	- -
ASP 72	- -	MET 70
MET 73	LYS 70	- -
VAL 74	ILE 71	- -
GLY 75	ILE 72	MET 90
THR 76	ALA 73	- -
PHE 77	GLU 74	- -
THR 78	- -	ALA 77
ASP 82	PRO 79	- -
PRO 83	- -	GLN 85
ALA 84	ALA 80	GLY 86
LYS 85	VAL 81	- -
PHE 86	PHE 82	- -

Table 4–4: C-α atoms in BLG and INSEC within 1.75 Å radius of a C-α atom in RBP. Table continued overleaf.

ASP 103	LYS 91	- -
HIS 104	VAL 92	- -
TRP 105	LEU 93	TRP 104
ILE 106	VAL 94	VAL 105
VAL 107	LEU 95	LEU 106
ASP 108	ASP 96	ALA 107
THR 109	- -	THR 108
ASP 110	ASP 98	ASP 109
TYR 111	TYR 99	TYR 110
ASP 112	LYS 100	LYS 111
THR 113	LYS 101	ASN 112
TYR 114	TYR 102	TYR 113
ALA 115	LEU 103	ALA 114
VAL 116	LEU 104	ILE 115
GLN 117	PHE 105	ASN 116
TYR 118	CYS 106	TYR 117
SER 119	MET 107	ASN 118
UNK 120	- -	CYS 119
ARG 121	ASN 109	ASP 120
LEU 122	- -	TYR 121
ASP 131	SER 116	SER 129
SER 132	LEU 117	ILE 130
TYR 133	VAL 118	HIS 131
SER 134	CYS 119	ALA 132
PHE 135	GLN 120	TRP 133
VAL 136	CYS 121	ILE 134
PHE 137	LEU 122	LEU 135
SER 138	VAL 123	SER 136
ARG 139	ARG 124	LYS 137
ASP 140	THR 125	SER 138
PRO 141	- -	LYS 139
GLY 143	- -	VAL 140
LEU 144	- -	LEU 141
PRO 145	- -	GLU 142
LEU 161	- -	ASP 161
ALA 162	LYS 141	- -
ARG 163	- -	SER 163
GLN 164	- -	LYS 164
TYR 165	- -	PHE 165
ARG 166	ILE 147	ILE 166
LEU 167	ARG 148	SER 167
ILE 168	- -	ASN 168

Figure 4–11: Structurally conserved α-carbon atoms for RBP, BLG, and INSEC. Structures superimposed using HYDRA, core calculated using a cutoff distance of 1.75 Å.

a2u		Source				
Residues	Sequence	File	Chain	Residues	Sequence	Rms (Å)
1	E	-			Random	-
2-5	EASS	BLG		1-4	LIVT	-
6-10	TSGNL	1PFC	A	24-28	PAINV	0.5527
40-49	VFMQHIDVLE	2ACT		168-177	GYGTEGGVDY	0.4286
49-55	ENSLGFK	RBP		54-60	SATAKGR	-
54-63	FKFRIKENGE	2CGA		9-19	LKMDKTKQPVV	0.2969
64-65	CR	RBP		70-71	CA	-
80-89	YFVEYDGGNT	1CPV		54-63	VWTRCNGGHW	0.2602
107-113	NFKNGET	1GP1		46-52	ERPNYQG	1.1344
119-126	LYGRTKDL	2MEV		258-265	YPTSGDKI	0.5676
126-137	LSSDIKEKFAKL	INSEC		141-154	LEGNTKEVVDNVLK	-
136-143	KLCEAHGI	1F19	H	62-69	EKFKGKTT	0.5084
143-153	ITRDNIIDLT	RBP		162-171	ARQYRLIVHN	-
151-157	LTKTDRC	1HDS	A	53-59	AQQKAHG	0.5690
157	C	RBP		174	C	-
158-162	LQARG	-			Random	-

Table 4–5: Fragments added to core lipocalycin structure to produce a full model of a2u. Rms fit for loop searches from SYBYL (version 5.4).

rapid movement towards a minimum far from the minimum. The GROMOS87 program PROEML was used with a long range cut-off of 15 Å and a short range cut-off of 8 Å. An initial step-size of 0.05 was used, the maximum step-size allowed was 0.1. The minimisation was terminated when the total energy change was less than 0.1 $kJmol^{-1}$. All bond lengths were constrained using the SHAKE algorithm, with a relative geometric tolerance of 0.001. The non-bonded interaction was calculated using an atom pair-list which was updated every 10 steps. After 200 cycles single point charge (SPC) model water molecules were added around the molecule to fill a truncated octahedron using the GROMOS87 PROBOX program. Steepest descents minimisation was then repeated until convergence. Minimisation in the presence of solvent was carried out on a Meiko Computing Surface using a parallel implementation of the GROMOS87 steepest descents minimiser. Finally, several cycles of conjugate gradients minimisation were performed using the serial version of PROEML. This final minimisation technique was used as it is a truely convergent method.

4.4.11 Validation of Model Structures

The above methods produced four different models for a2u, the next task was to determine which model was closest to the correct structure. The assessment of model correctness has been an area of much research (Novotny *et al.*, 1984; Novotny *et al.*, 1988; Baumann *et al.*, 1989; Hendlich *et al*, 1990) and is still problematic.

Overall Topology

Visual analysis of the four model structures indicated that they all have the overall topology of the lipocalycin fold (figure 4–12). The extent of β-strand and α-helix content in each structure, as calculated by the program DSSP (Kabsch and Sander, 1983), varies (table 4–6).

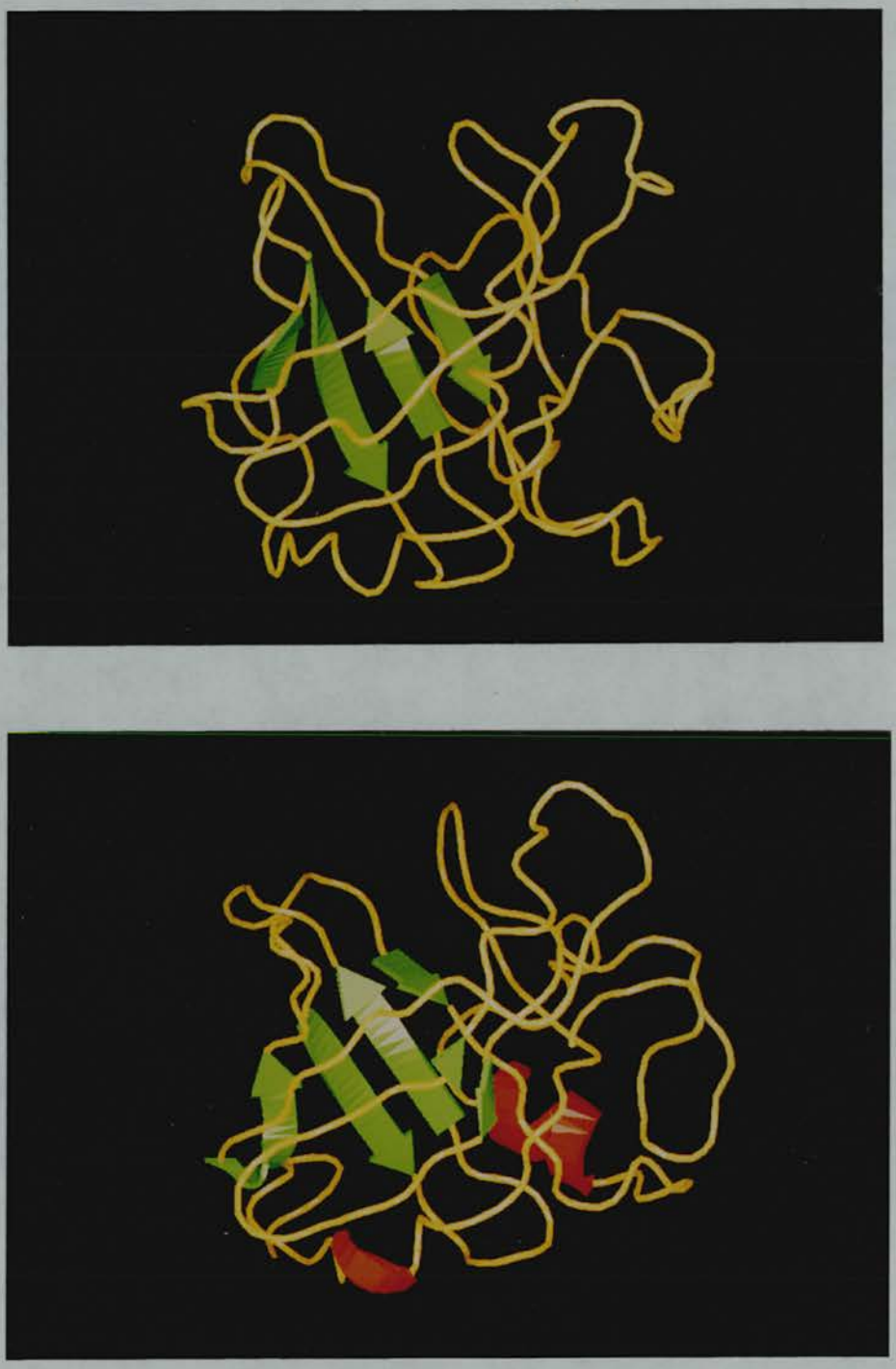
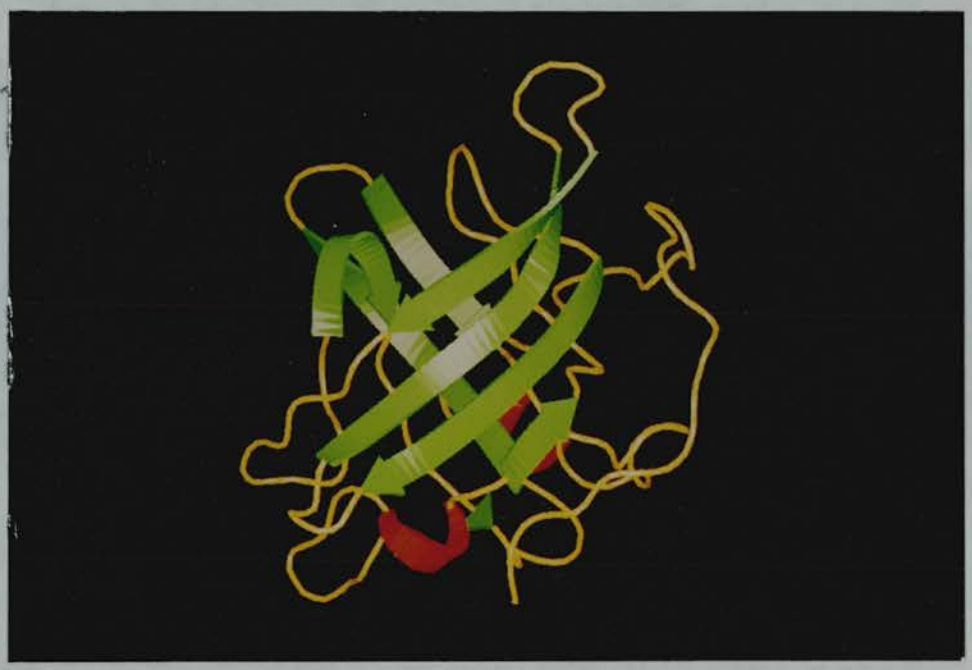
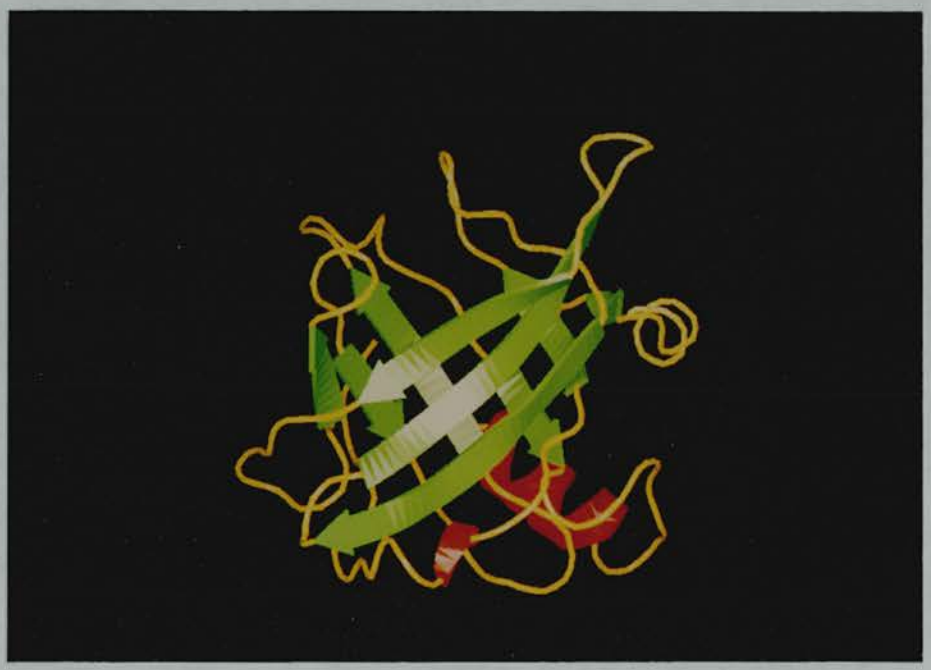


Figure 4–12: Cartoon representation of models 1 to 4. Model 1 (top), model 2 (below), continued overleafwith Model 3 (top), model 4 (below).





Model	1	2	3	4
β1 (A)	-	-	17-20	20-25
β2 (B)	-	37-39	34-43	35-44
β3 (C)	-	50-51	47-57	47-56
β4 (D)	-	-	63-70	63-69
β5 (E)	79-82	75-81	82-89	79-82
β6 (F)	88-94	89-94	94-95	88-91
β7 (G)	102-105	102-106	101-109	101-106
β8 (H)	116-119	115-117	112-117	116-120
α1	-	129-135	131-136	129-138
β9 (I)	-	-	-	146-148

Table 4–6: Secondary structure assignments for each model, calculated using DSSP.

Main-chain Geometry

Earlier, it was stated that the nature of the side chain groups in polypeptide chains places restrictions on the torsional angles between backbone atoms. The energetically favourable combination of peptide plane angles, ϕ and ψ, have been defined (Ramachandran and Sasisekharan, 1968; figure 4–2). Analysis of well refined, high resolution protein structures shows that these regions are indeed favoured (Thornton *et al.*, 1990). The distibution of mainchain torsional angles is therefore an indicator of protein-like structure. The distribution for glycine is not constrained, as it has no non-hydrogen side chain atom. Also proline has a somewhat different favoured distribution because of its topology. Bearing this in mind, we expect a native-like model structure to show a ϕ, ψ distribution conforming to the theoretical one. This does not mean that a good ϕ, ψ distribution indicates a native structure, but rather a native structure must have a favourable ϕ, ψ distribution. The mainchain torsional angles were calculated for each model using the program PHIPSI (PDB). The results are shown as conventional Ramachandran plots (figure 4–13). As reference the same plots were calculated for the structures used in the modelling protocols (figure 4–14). The ϕ, ψ distributions were analysed using the three observational sets defined by Thornton (table 4–7).

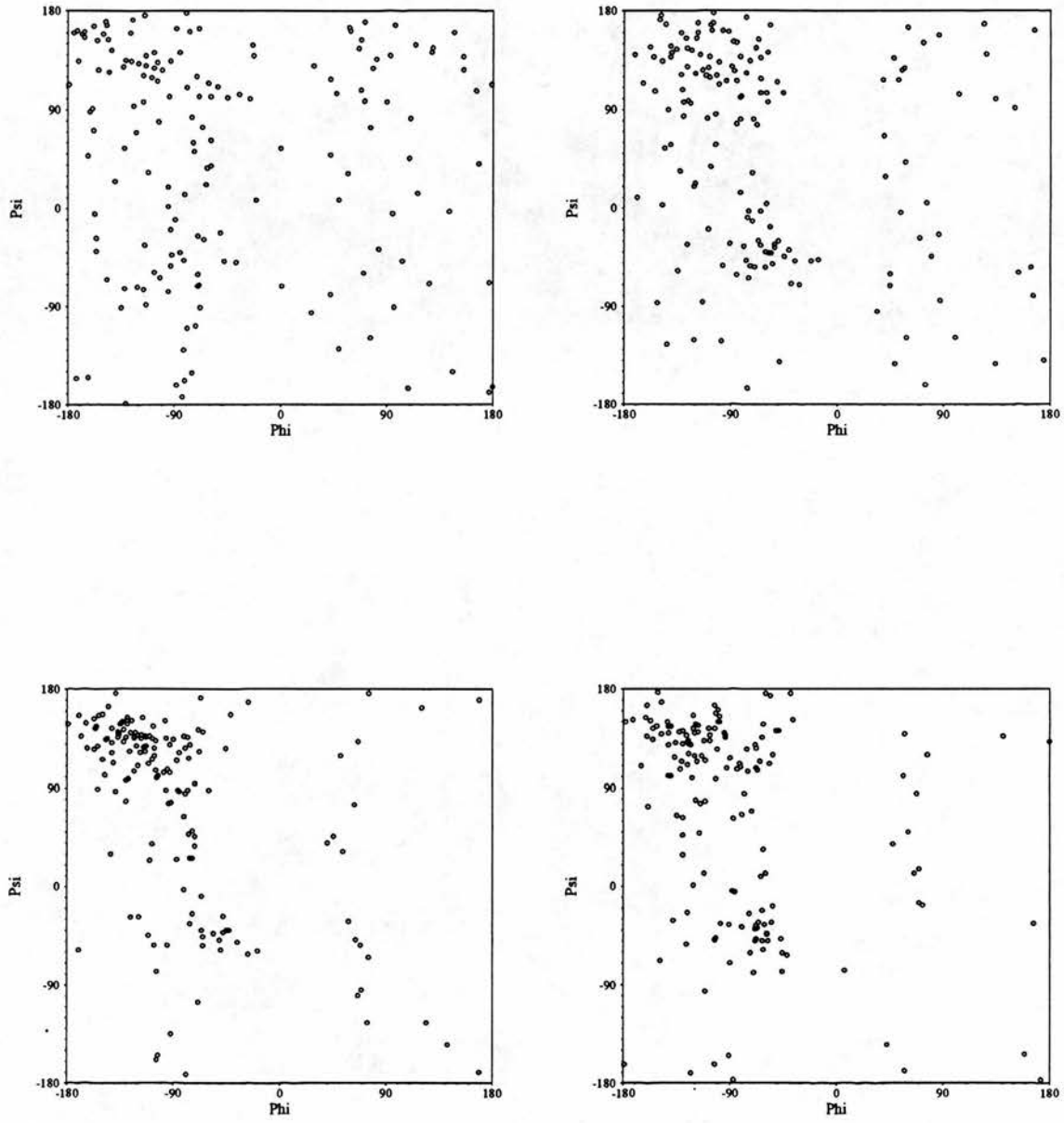


Figure 4–13: Ramachandran plots for model 1 (top left), model 2 (top right), model 3 (bottom left) and model 4 (bottom right).

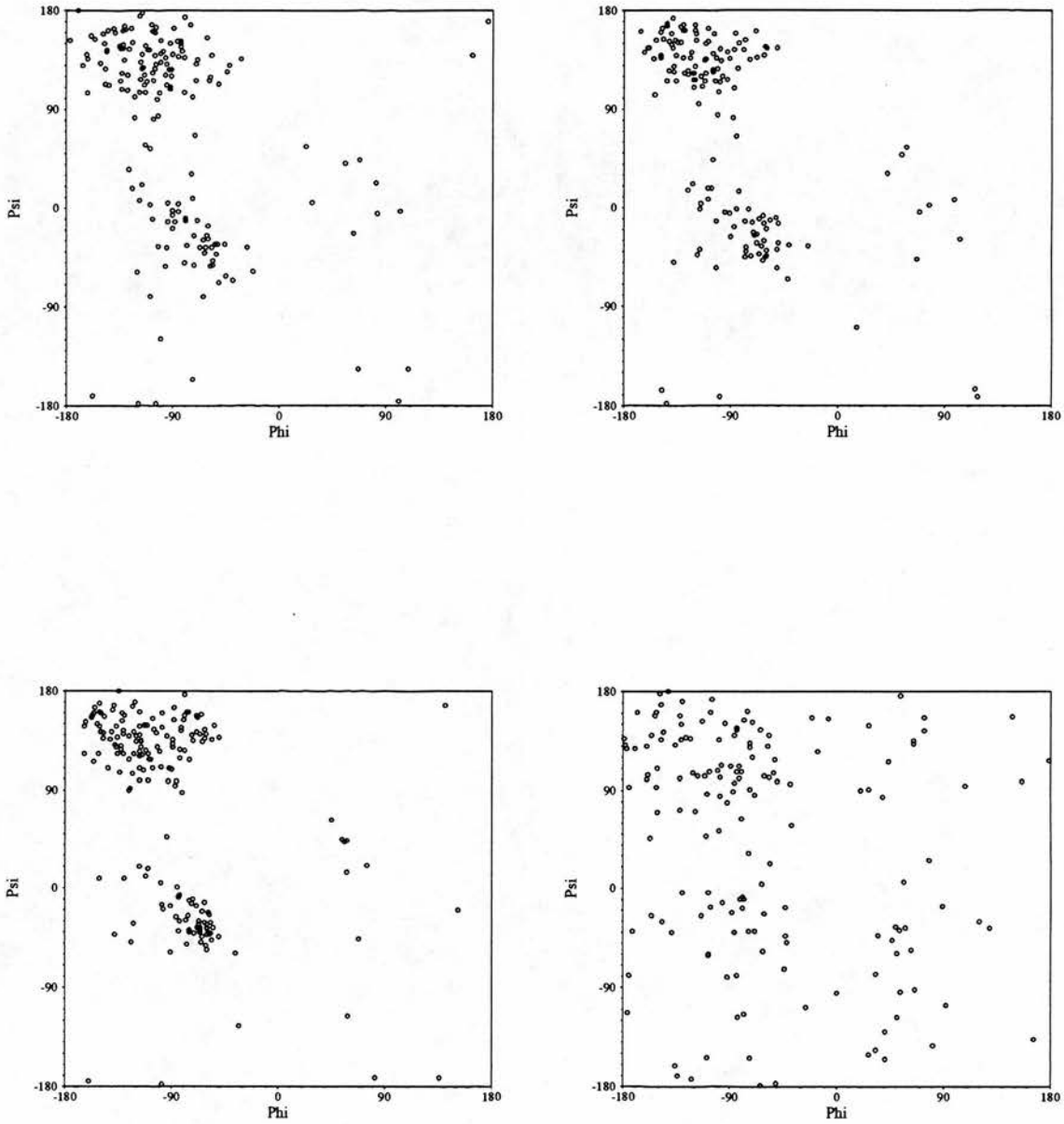


Figure 4–14: Ramachandran plots for RBP (top left), MUP (top right), INSEC (bottom left), and BLG (bottom right).

ϕ-ψ assignment	1	2	3	4
Allowed	114	129	149	153
Generous	25	14	3	8
Disallowed	23	19	10	1

Table 4–7: Semi-quantitative analysis of Ramachandran plots for models 1 to 4.

	1	2	3	4	1SA
Bonds (Å)	0.020	0.019	0.018	0.013	0.013
Angles (°)	4.172	4.055	3.124	3.675	4.528
Dihedrals (°)	29.476	27.074	27.495	29.323	35.678
Improper Dihedrals (°)	5.97	6.265	3.971	4.571	14.887

Table 4–8: Rms deviations of covalent parameters from ideality for models 1 to 4, and 1SA.

Deviations from Ideal Geometry

Mainchain torsional angles are generally constrained within defined limits. This also applies to the other covalent parameters of a protein molecule. The bond lengths between covalently linked atoms diverge little from their lowest energy states. The rms deviation seen in crystallographic structures is usually 0.02 Å and lower. Angles between three and four atoms are less rigidly defined as they can be deformed in order to satisfy both bond lengths and non-covalent interactions. Typically, angles between three atoms show an rms deviation of 3° from equilibrium positions. The rms deviations decrease as the resolution of the data increases and as refinement proceeds. As with the Ramachandran analysis native-like structures should show deviations in covalent geometry similar to that of crystal structures. The calculation of deviation of covalent parameters from ideal values was carried out with X-PLOR (Brünger, 1990) (table 4–8). This was done in a attempt to remove any possible bias introduced using one particular program for both minimisation and calculation of deviations from ideality. The forcefields for both GROMOS87 and X-PLOR are derived from the same initial parameters (Weiner *et al.*, 1984).

	1	2	3	4	1SA	RBP	INSEC	BLG
ANAREA	10479	10999	11183	10315	10358	9736	10659	10022
DSSP	11701	12146	11672	11019	11108	10259	11572	10303

Table 4–9: Calculated solvent accessible surface areas for models 1 to 4, 1SA, RBP, INSEC and BLG (in $Å^2$).

Solvent Accessible Surface Area

Empirical observation of crystallographic structures for monomeric proteins of between 50 and 320 amino acids has suggested a relationship between the solvent accessible surface area of these proteins and their molecular mass (Janin, 1976).

$$\text{Area} = 11.1 M_r^{\frac{2}{3}} \tag{4.42}$$

This relationship breaks down for oligomeric proteins whose monomers have 330 to 840 residues, in this case the surface area is approximately proportional to the molecular mass (Janin, 1979). This apparently anomalous observation has been rationalised by consideration of the area of amino acids that are buried upon folding (Janin and Chothia, 1979). Solvent accessible surface areas for model structures were calculated using the ANAREA program (Richmond, 1984). These were compared to the areas calculated for RBP, INSEC, and BLG (table 4–9).

Energy of Solvation

It is thought that of the forces that guide a protein to its final folded comformation, solvent interactions, including the hydrophobic force described earlier, are among the most important (Eisenberg and McLachlan, 1986). Different models for calculating the contribution of solvation energy to protein stability have been used. A method based on the exposed surface area for each atom in a structure and an atomic solvation parameter for each atom type has been used (Eisenberg and McLachlan, 1986). This method was implemented by the author to measure the relative solvation energy for each model. Energies of solvation were calculated using the solvent accessible area for each atom, as

	1	2	3	4	1SA	RBP	INSEC	BLG
GROMOS	-12708	-10092	-5015	-20138				
ANAREA	-44.1	26.4	-111.9	-51.1	-56.8	-30.1	7.3	78.5

Table 4–10: Calculated energies of solvation for models 1 to 4, 1SA, RBP, INSEC and BLG (in $kJmol^{-1}$).

calculated by the program ANAREA, and the atomic solvation parameters defined previously by Eisenberg and McLachlan. The results for the different models are compared to results obtained for RBP, INSEC and BLG (table 4–10).

Disulphide bridges

As shown in chapter 2 sequence alignments suggest a conserved disulphide bond in most members of the lipocalycin family, excluding INSEC, BBP and APOD. The sequence of a2u and its alignment with other lipocalycins suggest that residues 64 and 157 form a disulphide bond. It was only possible to model this bond in models one and four. The disulphide should contribute to structural stability but has not been demonstrated experimentally to exist in a2u.

4.4.12 Conclusions from Model Analysis

Considering intrinsic protein-like properties first. It can be seen that the ϕ, ψ distribution for the model structures improves from model one to four (figure 4–13). This very poor distribution for model one is attributed to a similarly bad distribution in the BLG structure used to model a2u in that method (figure 4–14). Attempts to improve model one will be discussed shortly. The bad Ramachandran plot seen for model two is also due to the large amount of BLG structure used in the construction of the modelling template (table 4–3). In contrast the ϕ, ψ distributions for models three and four are within the limits seen for refined crystal structures. Classification of ϕ, ψ pairs using the three groups suggested by Thornton (Thornton *et al.*, 1990) provides a semi-quantitative analysis of the distributions (table 4–7). The results suggest that model four has the most favourable mainchain torsional angle distribution.

The deviations from ideal covalent parameters also show a minimum for model four (table 4–8). The deviation from ideality does not vary much between the four models, even for the models based on the BLG structure. This observation is perhaps not surprising, as the prolonged minimisation carried out on the models will have moved the atoms such that covalent parameters are close to their optimum. The larger rms deviations for some models is presumably due to non-bonded interactions between incorrectly placed atoms. Energy minimsation may optimise the position of these residues at the expense of the covalent geometry.

The accessible surface areas calculated by both ANAREA and DSSP indicated that model 4 had the lowest surface area. In both cases the area was larger than that expected (approximately 8000 Å^2). Analysis of the surface areas of RBP, INSEC and BLG suggested that they do not obey eqn 4.42. Comparison of the expected values and those observed above indicated that the areas calculated directly from the structure were approximately 1.3 times greater than expected. The modified surface area equation becomes,

$$\text{Area} = 14.4 M_r^{\frac{2}{3}} \tag{4.43}$$

This apparent underestimation of surface area by eqn 4.42 is probably because the ligands were not included in the direct calculation of the surface area. The lipocalycin structures solved possess deep binding pockets. If the crystallographic ligand is removed there is a large internal area exposed to the solvent - which is included in the surface area calculated by either ANAREA or DSSP. The surface area removed from possible interaction with the solvent upon binding of retinol is 251 Å^2 for RBP (Cowan *et al.*, 1990). The expected surface area of a2u was therefore calculated assuming that the structure modelled was that of the protein/ligand complex with the ligand removed. The expected surface area was therefore approximately 9500 Å^2. The observed surface areas were still higher than the revised estimate, suggesting that the modelling procedures had all failed to produce a compact structure. This may have been to

the inherent limitations of energy minimisation algorithms, or due to the incorrectness of the models. Loop regions are difficult to model correctly and also make a large contribution to the surface area of the molecule.

The calculated energies of solvation for the 4 models show great variation (table 4–10). The figures derived from the final energy minimisation results all indicate a negative potential energy interaction between solvent and protein. However, it should be remembered that this is a 'static' figure, representing only one solvent/protein configuration. The values calculated from solvent accessible surface area results make no assumption about the position of the solvent. These values should really be considered as mean solvation energies. The algorithm assumes that the solvation energy is proportional to the exposed surface area, but takes no account of the allowed packing of solvent molecules around the protein. Both methods have their deficiences and the results therefore must be treated with some caution. The lowest potential energy between solvent and protein from energy minimisation was model 4, while the lowest from surface area calulations was model 3. The contradictary results for model 3 cannot be readily explained. One possibility is that the atomic solvation parameters used require revision. The values derived from surface areas were within the same region as those for RBP and INSEC. Considering both sets of figures together suggested that model 4 was a model with acceptable protein/solvent interactions and a calculated solvation energy similar to that observed for RBP.

It was only possible to model the expected disulphide bond between residues 64 and 157 in methods 1 and 4. This fact alone suggested that models 2 and 3 could not be candidates for the correct a2u structure. The disulphide bonds in model 1 and model 4 were in similar positions when the structures were superimposed. However the main chain conformation around the region of the disulphide bond in model 1 was markedly different to that of model 4. A Ramachandran plot of this region in both molecules showed that model 1 had very poor mainchain geometry (figure 4–15).

The analyses carried out suggested that the most correct model was model 4. However, to verify this a comparison to the crystal structure of a2u needed to be

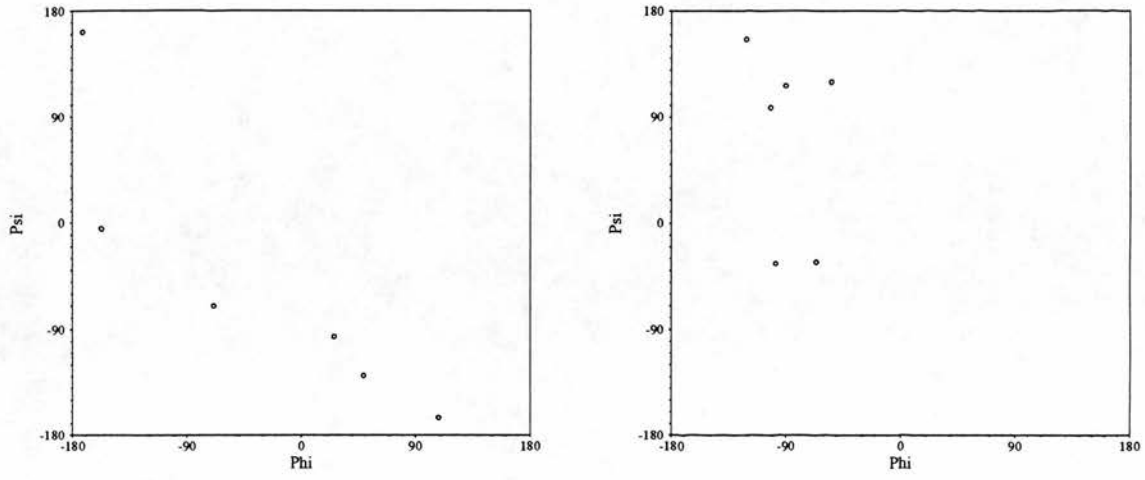


Figure 4–15: Ramachandran plot for residues 63 to 65 and 156 to 158 in model 1 (left) and model 4 (right).

made. This would have been possible had problems with the crystal form of a2u not been encountered (chapter 3). Fortunately, the crystal structure of MUP was solved by other workers and the coordinates made available for comparison. First the problems with the structure of BLG which resulted in model 1 are discussed, followed by description of an attempt to rectify some of the problems with this model.

4.4.13 Why BLG was a Bad Modelling Template

Analysis of the structure of BLG after its use for modelling indicated that it was unsuitable for such a task. The Ramachandran plot showed that many residues had un-protein-like main-chain dihedral angles (figure 4–14). Analysis of the structure with the DSSP program indicated very little secondary structure (20 residues in a β-strand conformation, and no residues in an α-helical conformation). Difficulties in refinement of the structure of BLG with either

conventional restrained least-squares, or molecular dynamics simulated annealing may be indicative of misplacement of some residues far from their correct position (A.S.McAlpine, personal communication). The sequence of feline lactoglobulin (FLG) was published recently (D.Shaw *et al*, 1991). When compared to the sequence of bovine lactoglobulin (BLG), using the GAP program from UWGCG (figure 4–16), a high level of identity was observed (54%). However, Gln35 in BLG is replaced by a Glu35 in FLG. In the present structure of BLG has the sidechain of Gln35 postioned in the hydrophobic calyx, and thus totally inaccessible to solvent. It is unlikely that a glutamate residue would be situated in such an environment, it being very energetically unfavourable to bury a charged group within a hydrophobic environment. However, the presence of carboxylic acid group with anomalous behaviour, possibly due to a hydrophobic environment, has been inferred from titration studies (Tanford *et al.*, 1959). The BLG sequence and structure were compared to those of MUP (figure 4–17 and figure 4–18). The structural superposition showed a surprising degree of similarity between the two structures (rms fit of 1.863 Å for 117 α-carbon atoms). However, the major difference observed was in the large loop between strand A and B. The loop is longer in MUP and therefore after Gly36 the sequence alignment and structural aligment are mismatched by 3 residues. The sequence alignment placed Ala37 in BLG next to Gly36 in MUP, the structure alignment had Arg40 in BLG close to Gly36. Residue Arg40 is well conserved amongst MUP, BLG, FLG and a2u suggesting a possible structural similarity in this region. The position, length and structure of strands B and C were very similar in BLG and MUP, but misaligned by 3 residues. The loop between strands C and D is two residues longer in BLG therefore the mismatch along strand D is only one residue. The loop between strand D and E is one residue longer in BLG thus returning the structural alignment to that of the sequence.

These observations suggest that the structure of BLG may be incorrect in the region between strand A and strand E, as a result of the loop between strand A and B being too short. Lengthening of this loop in BLG by three residues would

```
          .         .         .         .         .
  1 LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTP 50
    .:..||.:|||..|||||.|:||||||||||||.:.|||||||:||:|||
  1 ATLPPTMEDLDIRQVAGTWHSMAMAASDISLLDSETAPLRVYVQELRPTP 50
          .         .         .         .         .
 51 EGDLEILLQKWENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYK 100
    ::|||:|.|:|| .| : .|:|::|. ||||.:|  .|.|: ||||||.
 51 RDNLEIILRKRENHACIEGNIMAQRTEDPAVFMVDYQGEKKISVLDTDYT 100
          .         .         .         .         .
101 KYLLFCMENSA.EPEQSLACQCLVRTPEVDDEALEKFDKALKALPMHIRL 149
    .|::||||..| :.|.:: ||:|.|| ..|:|.:||||:||..||:|||:
101 HYMFFCMEAPAPGTENGMMCQYLARTLKADNEVMEKFDRALQTLPVHIRI 150
          .
150 SFNPTQLEEQCHI 162
     :: || .|||::
151 ILDLTQGKEQCRV 163
```

Figure 4–16: Sequences of bovine and feline lactoglobulin aligned with the program GAP using default parameters.

```
> blg           LIVTQTMKGLDIQKVAGTWYSLAMAASDISLLDAQSAPLRVYVEELKPTPEGDLEILLQK
> catblg        ATLPPTMEDLDIRQVAGTWHSMAMAASDISLLDSETAPLRVYVQELRPTPRDNLEIILRK
> mup           EEASSTGRNFNVEKINGEWHTIILASDKREKIE-DNGNFRLFLEQIH-VLENSLVLKFHT
> a2u           EEASSTRGNLDVAKLNGDWFSIVVASNKREKIE-ENGSMRVFMQHID-VLENSLGFKFRI
                  . *  .... ...*.* .. .*... . .. ... .*......  .  ..* . ..

> blg           WENGECAQKKIIAEKTKIPAVFKIDALNENKVLVLDTDYKKYLLFCMENSA-EPEQSLAC
> catblg        RENHACIEGNIMAQRTEDPAVFMVDYQGEKKISVLDTDYTHYMFFCMEAPAPGTENGMMC
> mup           VRDEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMG
> a2u           KENGECRELYLVAYKTPEDGEYFVEYDGGNTFTILKTDYDRYVMFHLINFKNGETFQLMV
                  . .* .  ..* .*   . . ..  . ..  . .***.....  . .   . .  .

> blg           QCLVRTPEVDDEALEKFDKALKALPMHIRLSFNPTQLEE--QCHI
> catblg        QYLARTLKADNEVMEKFDRALQTLPVHIRIILDLTQGKE--QCRV
> mup           LY-GREPDLSSDIKERFAQLCEEHGILRENIIDLSNANRCLQARE
> a2u           LY-GRTKDLSSDIKEKFAKLCEAHGITRDNIIDLTKTDRCLQARG
                 .  *. . ...  *.*..  ..  .    .. .. .  * .
```

Figure 4–17: Sequences of BLG, FLG, a2u, and MUP aligned with the program CLUSTAL using default parameters.

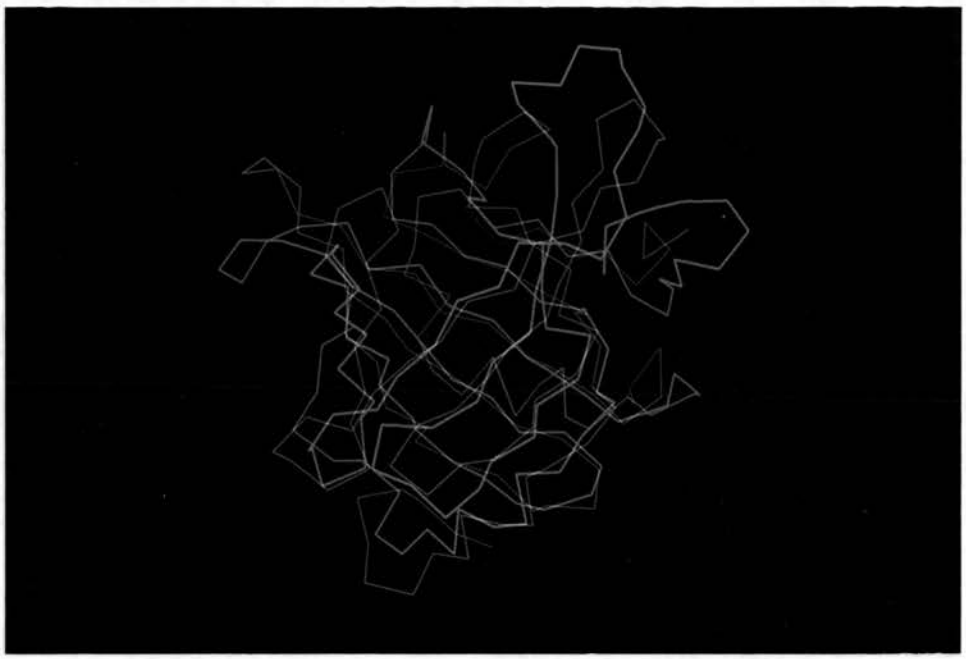


Figure 4–18: Structures of BLG and MUP superimposed using the program O.

bring Gln/Glu35 into a position accessible to solvent (within the loop itself). Attempts to redetermine the structure of BLG by *ab initio* model building with reference to the MUP stucture, and molecular replacement using MUP as a search molecule are in progress (A.S.McAlpine, personal communication).

4.4.14 Simulated Annealing of Model 1

An attempt to improve model one was made by using a simulated annealing technique. Annealing is a physical process in which a solid in a heat bath is heated to a point at which the particles in the solid randomly arrange themselves in a liquid-like phase, followed by cooling slowly by lowering the temperature of the surrounding heat bath. The particles arrange themselves in the lowest energy ground state of the solid, provided the heating has been to a high enough temperature and the cooling is slow enough. The term simulated annealing (SA) is used to refer to the application of the annealing process to optimisation problems. The target of the optimisation problem is identified with the energy of the system (Kirkpatrick, 1983). The advantage of SA over other optimisation methods is its ability to overcome energy barriers by searching in an uphill

direction. The likelihood of overcoming barriers is related to the temperature - the likelihood increases as the temperature increases. Temperature can be included explicitly in molecular dynamics calculations (Berendsen *et al.*, 1984). It is therefore possible to carry out an SA procedure with a protein structure using simulation at elevated temperature followed by slow cooling. This technique has become central to the refinement of both crystal and NMR structures (Petsko and Karplus, 1991). Simulated annealing was used to search for a better energy minimum for the poor model produced by method 1. The atoms for model plus 2923 surrounding SPC water molecules were assigned initial velocities from a Maxwellian distribution at 600 K. The system was then simulated at constant pressure for 2 ps at 1 fs timesteps, at which point an equilibrium had been reached. The system was then cooled to 310 K over 500 steps of 1 fs. The temperature was controlled by coupling to an external heat bath using a coupling constant of 0.01 ps (Berendsen *et al.*, 1984). A further 500 fs of simulation were carried out, the coordinates were stored every 20 fs and averaged to give a final structure (1SA). The SHAKE algorithm was used to constrain all bond lengths during simulation. The Ramachandran plot after annealing (figure 4–19) is slightly improved compared to the initial model but still has many residues in disallowed conformations. The surface area (table 4–9) is reduced and the energy of solvation increased (table 4–10). However, the geometry of the structure becomes distorted, with increased deviations from ideality for angles, proper dihedral, and improper dihedral angles (table 4–8). The deviation in bond length is improved after simulated annealing.

4.4.15 Comparison between Model 4 and MUP

It is assumed that a2u and MUP possess essentially the same structure because of the high level of identity between their sequences (66%). The crystallographic structure of MUP was solved in 1991 (Böcskei *et al.*, 1991; Flower, 1992). The coordinates were kindly made available by A. C. T. North for the purpose of analysing the different models of a2u. The coordinates provided had been refined to an R-factor of approximately 27% at 2.4 Å resolution using X-PLOR. The

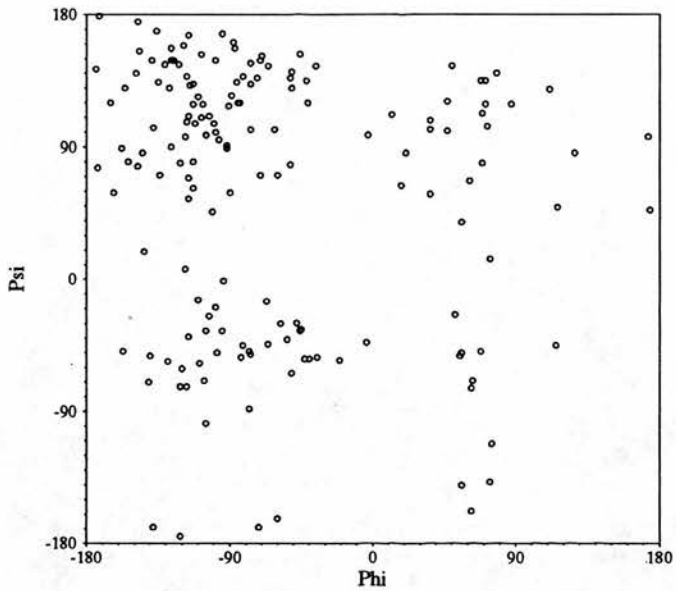


Figure 4–19: Ramachandran plot for model one after simulated annealing.

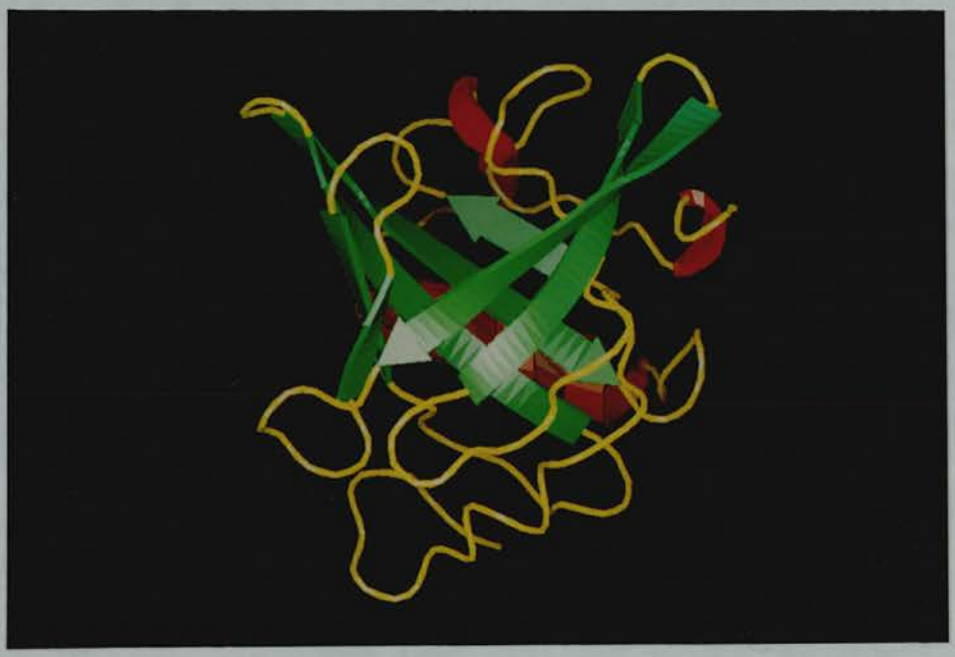


Figure 4–20: Cartoon representation of MUP.

Bonds (Å)	Angles (°)	Proper Dihedrals (°)	Improper Dihedrals (°)
0.025	5.568	32.196	2.547

Table 4–11: Rms deviations from ideal geometry for MUP.

model possesses the lipocalycin β-barrel fold (figure 4–20) and shows geometry consistant with a native protein structure, as assessed by the Ramachandran plot (figure 4–14). The rms deviations from ideal geometry (table 4–11) are high compared to other refined structures, this is probably because the coordinates supplied were not those of the final fully refined model. The crystallographic model comprises residues 2 to 157 as electron density is poor for the N and C terminii. The structures of the different models (residues 2 to 157 in both) were superimposed on to MUP using the lsq_explicit and lsq_improve commands in the program O (figure 4–21). The rms fit between each model and MUP for the number of matched residues found by the lsq_improve algorithm is presented as an indication of structural similarity (table 4–12).

1	2	3	4
7.615	6.767	7.039	6.369

Table 4–12: Overall rms deviations between matched α-carbon atoms for models 1 to 4 and MUP.

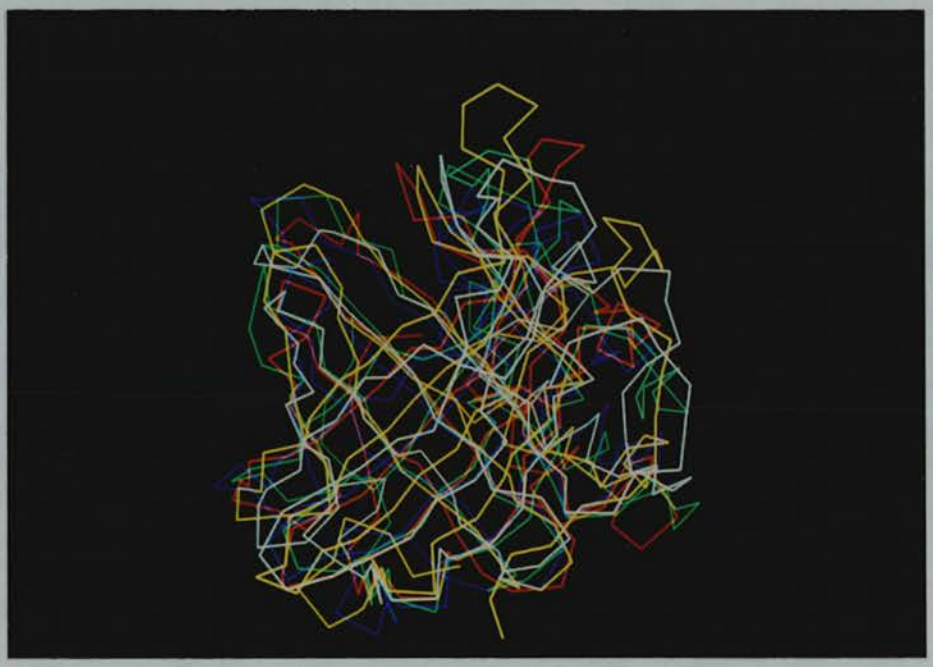


Figure 4–21: Superposition of model 1 (green), 2 (white), 3 (yellow), 4 (red) and MUP (blue).

In order to make a comparison between model 4 and the structure of a2u, a model of a2u was constructed based on the X-ray model of MUP (a2umup). This was achieved by mutation of the residues of MUP to a2u within SYBYL. The resulting structure was energy minimised *in vacuo* using X-PLOR. A limited number of steps was used - 200 cycles of congugate gradients minimisation. Also, α-carbon atoms were harmonically restrained (20 kcal mol^{-1} $Å^2$). This was done to restrain the overall structure of the model to that of MUP. The minimisation was primarily carried out to remove bad steric interactions between the mutated side-chains, not to optimise the complete structure. It was felt that this approach was valid because of the high level of sequence identity between a2u and MUP - suggesting the deviation between their structures is small. The model derived from MUP (a2umup) was superimposed on model 4 using lsq_explicit and lsq_implicit in the program O (figure 4–23). The rms difference in α-carbon position between equivalent residues in the sequence was calculated using X-PLOR (figure 4–22). This graphically indicated where large deviations between the two structures occur. The major differences were at the N-terminus, the loops and strands between residues 35 and 60, and from the α-helix to the

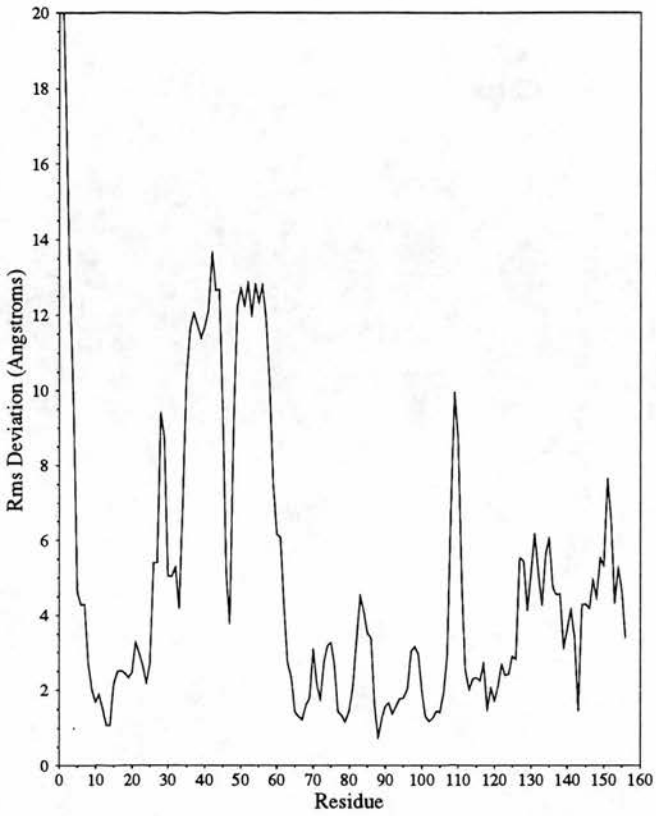


Figure 4–22: Rms deviations between the α-carbons of equivalent residues in a2umup and model 4.

C-terminus.The secondary structure for both a2umup and model 4 was calculated using the program DSSP. The models were compared on the basis of the secondary structure elements seen in a2umup. Only those residues observed in the crystal structure of MUP (2 to 157) were considered. In each section, analysis of a2u modelled from MUP is presented first followed by analysis of model 4.

Residues 2 to 16

The N-terminal residue, Glu1 was not resolved in the electron density map for MUP. Residues Glu2 to Ser4 form a linear chain which runs parallel to strand F. Residues Ser5 and Thr6 form a bend which becomes a β-turn stabilised by

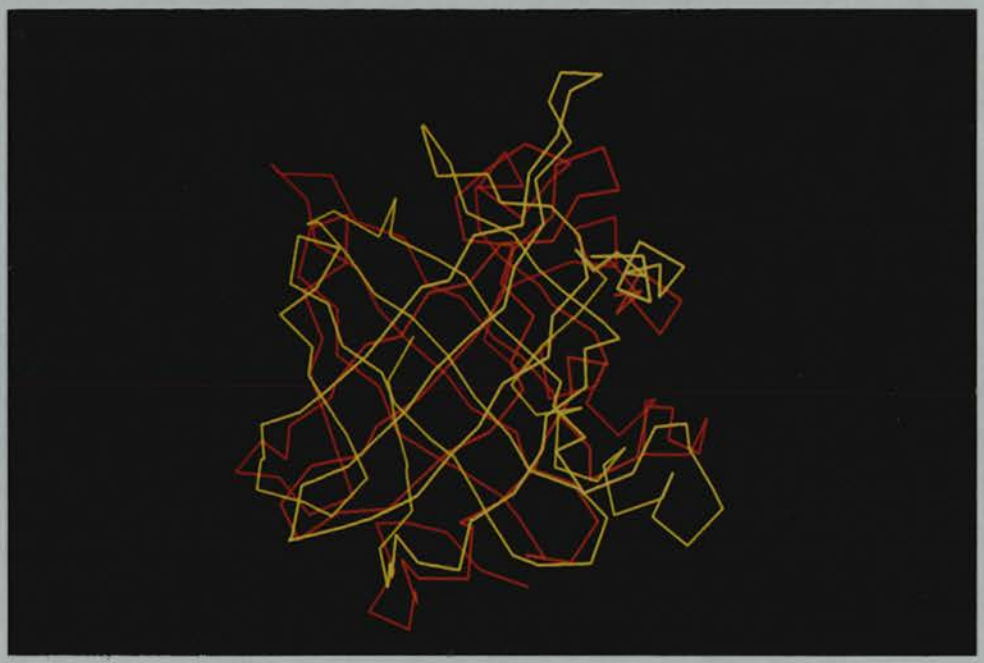


Figure 4–23: Model 4 (yellow) and a2umup (red) superimposed using the program O.

hydrogen bonding between Ser7 and Leu10. Residues Asp11 to Asn16 form another β-turn with hydrogen bonds from residue Asp11 to Lys14, and Ala13 to Asn16.

Model 4 has a less regular N-terminal region. The first residues do not run parallel to strand F, instead they point out into the solvent. A bend centered on Asn9 is the first item of secondary structure. Residues Val12 to Asn16 form a 3_{10} helix with hydrogen bonds from residues Val12 to Leu15, and Ala13 to Asn16. These are the same residues which form a β-turn in a2umup.

Residues 2 to 16 from both a2umup and model 4 are shown in figure 4–24.

Residues 17 to 26

Residues Gly17 and Asp18 form a β-strand, hydrogen bonding with Ile45 and His44 respectively. Residue Gly17 is the glycine in the conserved G-T-W motif and as such its interaction with other residues is expected to be of importance. Residue Asp18 forms a hydrogen bond from OD1 to Glu43 O. Residue Trp19 forms a hydrogen bond with Gln43 but does not conform to the β-strand

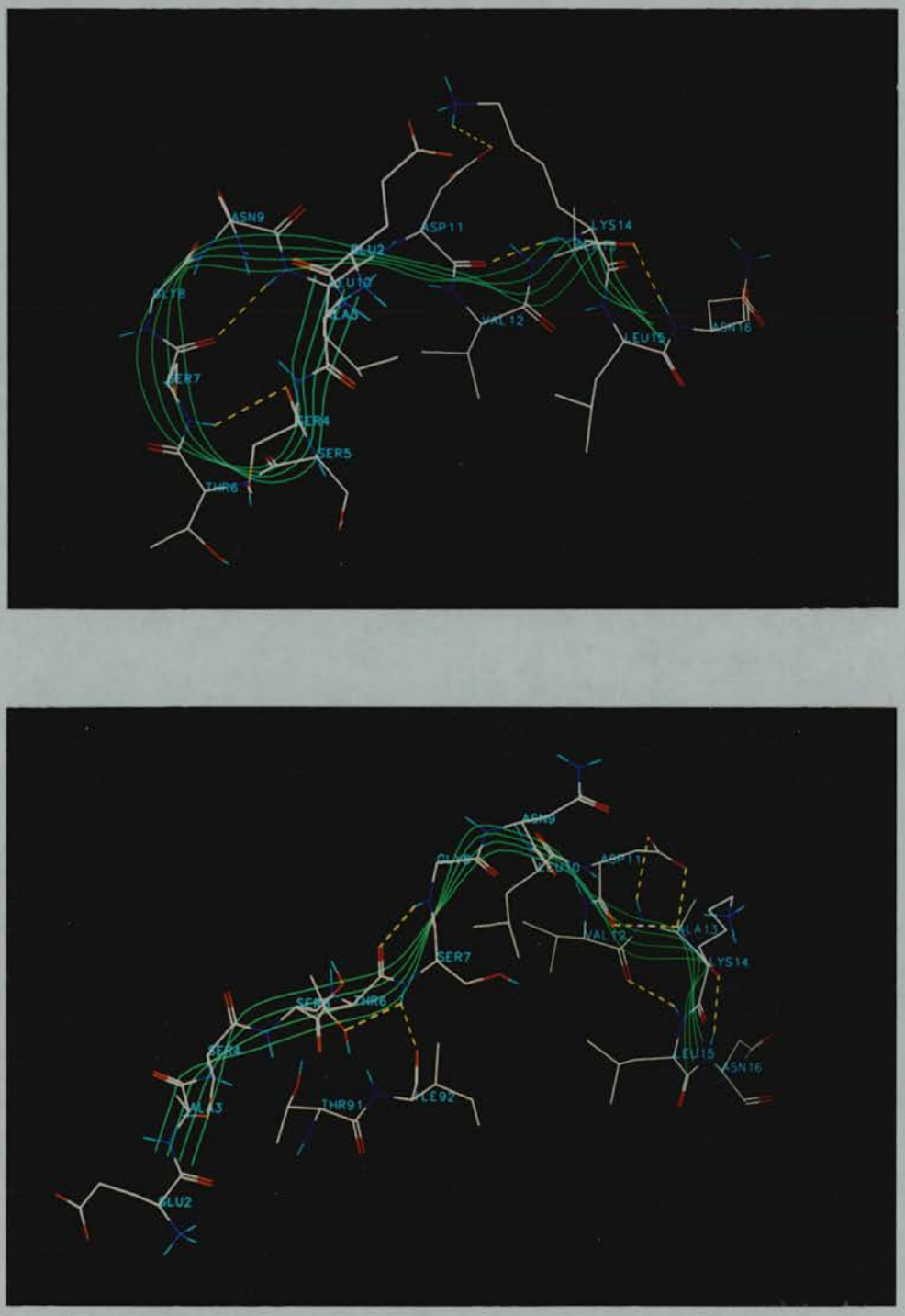


Figure 4–24: Residues 2 to 16 from a2umup (top) and model 4 (bottom).

topology. The hydrophobic region of the Arg122 sidechain packs against the face of Trp19. A similar sort of packing is seen between the alkyl chain of Lys124 and Phe20. Residue Ser21 forms a hydrogen bond from its sidechain oxygen (OG1) to the mainchain nitrogen of Met42. Residues Ser21 and Ile22 form a β-bulge changing the chain direction, hydrogen bonds are to Gly121 and Tyr120 respectively (strand H). This bulge is part of a longer β-strand running from Phe20 to Ser26, which is strand A in the lipocalycin topology. Residues Val23 to Ala25 form hydrogen bonds with Leu119 to Met117 (strand H) and Asp150 to Ile148 (strand I). Residue Ser26 continues the interaction with strand H by hydrogen bonding to Leu116.

In the model 4 Gly17 forms a β-bridge to residue Phe41. No packing of Arg122 across the face of Trp19 is seen. Arg122 is infact situated approximately 8 Å away from the face of Trp19. The deviation between mainchain positions of Trp19 for the two structures is less than 1 Å, but the sidechain position is very different in model 4. The ring is rotated 180° away from its position in a2umup, by a simple rotation around the Cα-Cβ bond, and a small shift in the mainchain position. Residues Ser21 and Ile22 form a β-bulge, hydrogen bonds being to Tyr120 and Leu119 respectively (strand H), and Met38 (strand B). They form part of the longer β-strand from Phe20 to Ala25 (strand A). Residues Val23 to Ala25 form hydrogen bonds with Ile148 to Asp146 (strand I), and Val118 to Leu116 (strand H) respectively.

Residues 17 to 26 from both a2umup and model 4 are shown in figure 4–25.

Residues 27 to 40

Residues Asn27 and Lys28 form a bend which then becomes a 3_{10} helix (residues Arg29 to Ile32). This helix is formed by hydrogen bonding from Lys28 to Lys31, Arg29 to Ile32, and Glu30 to Glu33. Residue Glu33 also participates in a β-turn identified at residues Glu34 and Asn35, which is stabilised by a hydrogen bond between Glu33 and Gly36. This is followed by another β-turn at residues Ser37

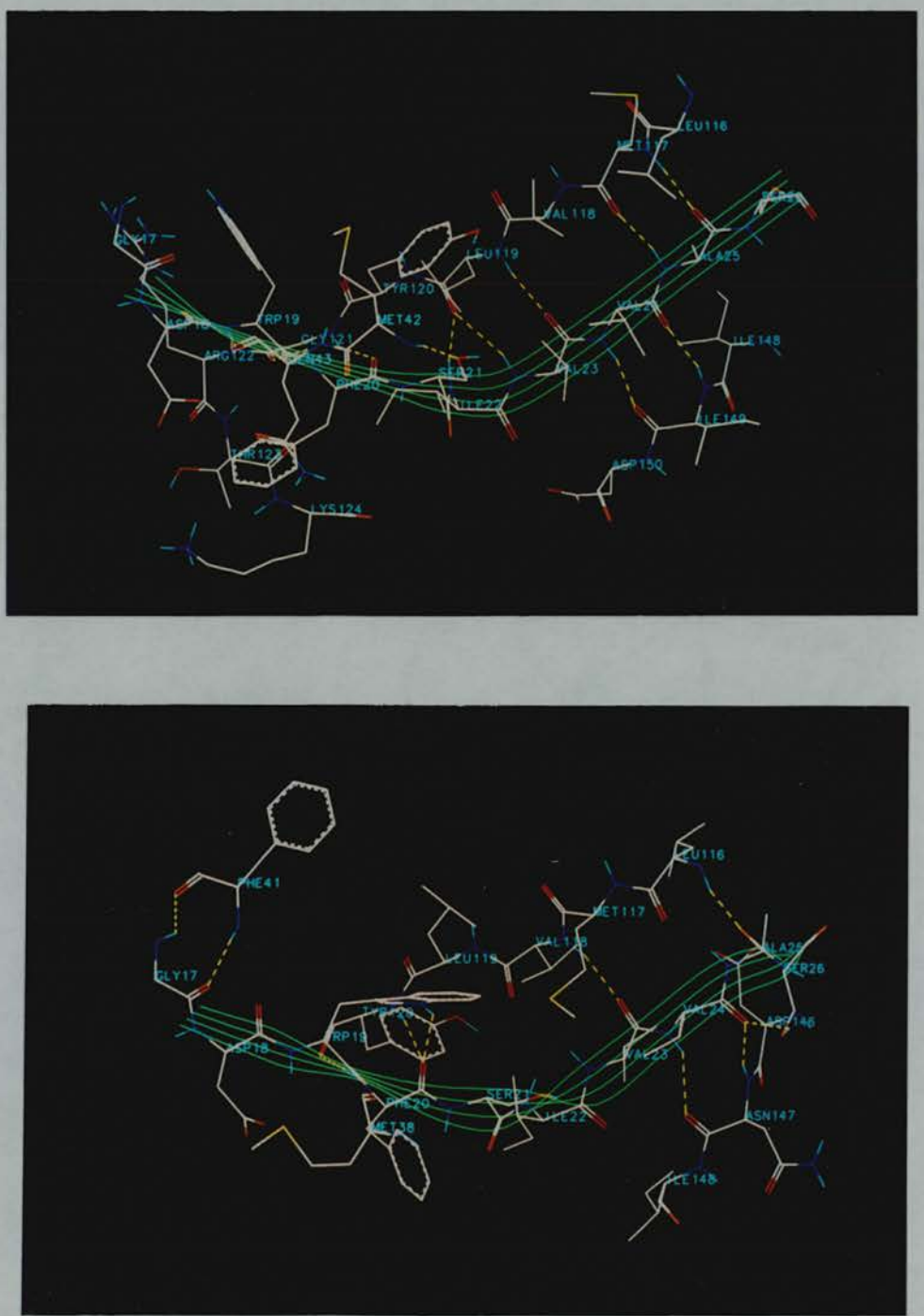


Figure 4–25: Residues 17 to 26 from a2umup (top) and model 4 (bottom).

and Met38, which is stabilised by hydrogen bonding between Gly36 and Arg39. The hydrophobic sidechain of Val40 stacks against the side of the ring of Phe56.

This region possesses less secondary structure in model 4, it is essentially a large open loop. The loop is defined by a β-turn at residues Arg29 and Glu30, which is stabilised by hydrogen bonding between Lys28 and Lys31. This large loop is shorter in length than that observed in a2umup (by some 4 amino acids). Consequently, the next β-strand in model 4 begins prematurely at residue Asn35. This β-strand (strand B) extends through to residue His44. Another consequence of the shorter loop in model 4 is the location of some residues in unexpected enviroments. Residue Ile32 is exposed to solvent, whilst it is situated in a hydrophobic environment in a2umup (lying close to Val24 and Ile149). Residues Glu33, Ser37, and Arg39 are located inside the mainly hydrophobic calyx, whereas they are exposed to solvent (as part of the large loop) in a2umup. It is unlikely that such polar residues would exist in a hydrophobic environment without forming salt-bridges, or hydrogen bonds - these are not observed in model 4.

Residues 27 to 40 from both a2umup and model 4 are shown in figure 4–26.

Residues 41 to 62

Residues Phe41 to Val47 form a β-strand (strand B). Residues His44 to Ile45 form hydrogen bonds with residue Gly17 (the G-x-W motif). Residues Phe41 to Met42 hydrogen bond with Arg57 to Phe56 (strand C), and Asp46 to Val47 hydrogen bond with Gly53 to Leu52 (strand C). Hydrophobic residues lie mainly on the interior face of the β-strands, with polar residues on the other face - exposed to the solvent. An exception is Leu48 which is situated on the loop between strands C and D - leaving it exposed to the solvent. Residues Glu49 and Asn50 form a bend which is followed by strand C starting at residue Ser51, continuing through to residue Glu60. This is a long β-strand (10 residues) which interacts with the previous strand B and the following strand D. Residues Leu52 to Arg57 hydrogen bond with Val47 to Phe41 (strand B) and Ala71 to Glu66

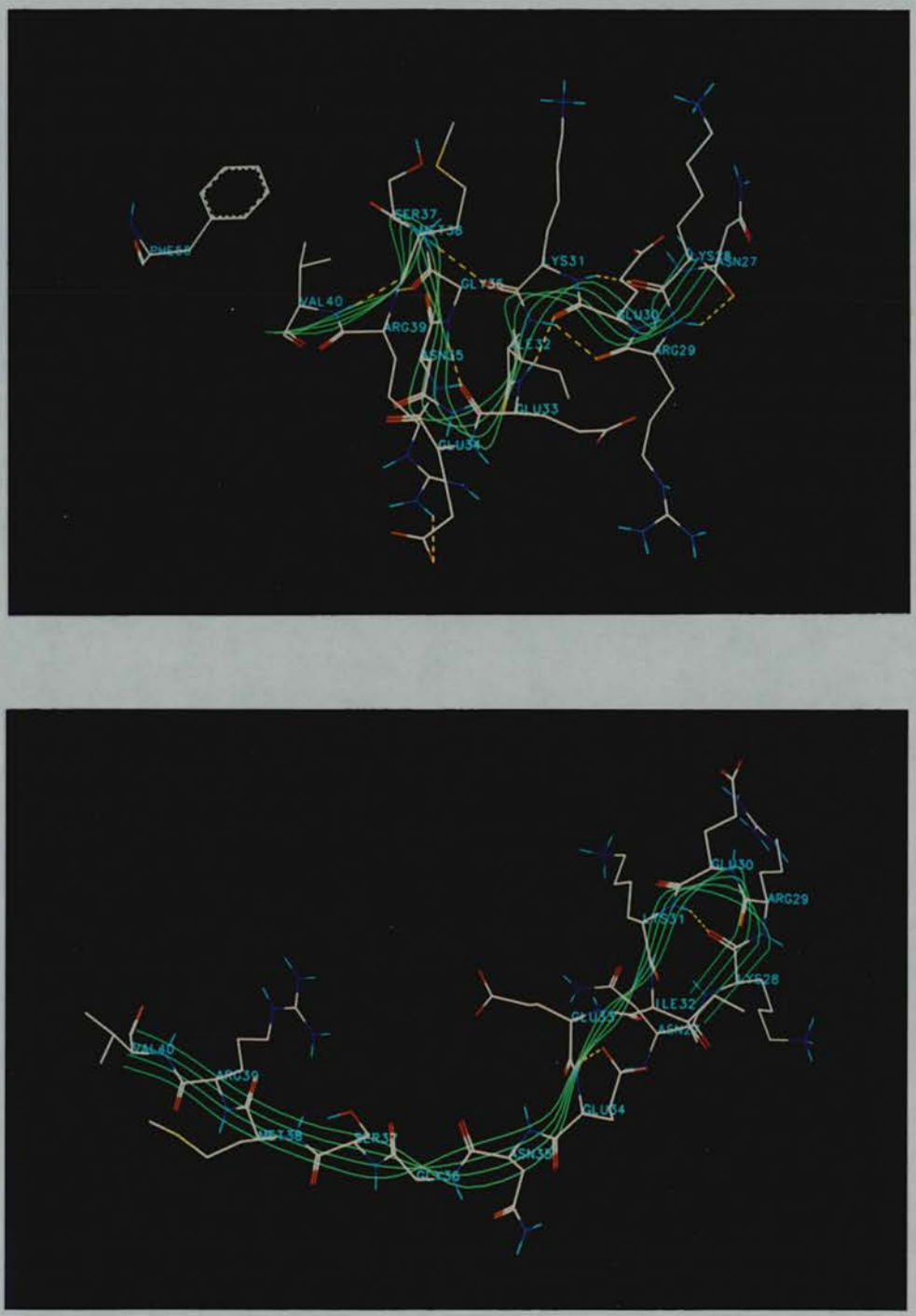


Figure 4–26: Residues 27 to 40 from a2umup (top) and model 4 (bottom).

(strand D). A continuous interaction is observed between strand C and D, with hydrogen bonding between Ser51 to Glu60 and Tyr72 to Glu63. This long β-strand is followed by a tight β-turn at residues Asn61 and Gly62, which is stabilised by a hydrogen bond between Glu60 and Glu63. This β-turn then leads into strand D.

Residue Phe41 is midway along a 10 residue β-strand (strand B) which begins at residue Asn35. Residues Ser37 to His44 form hydrogen bonds with Phe54 to Val47 (strand C). Residue Asn35 forms a hydrogen bond with Lys55, while Phe41 interacts with Gly17. The strand is terminated by a tight β-turn at Ile45 and Asp46, which is stabilised by a hydrogen bond between His44 and Val47. This turn then leads into another long β-strand (strand C) formed by residues Val47 to Phe56. Residues Asn50 to Lys55 form hydrogen bonds with Phe41 to Ile45 (strand B) and Leu69 to Cys64 (strand D). Residues Val47 to Glu49 hydrogen bond with His44 to Met42 (strand B), while Phe56 forms a hydrogen bond with Glu63 (strand D). This strand is terminated by an ill-defined loop stretching from residues Arg57 to Gly62. The premature start of the strand C results in Ile45 and Ile47 remaining exposed in the loop between C and D. Residue Asn50 is situated inside the protein in the hydrophobic calyx. Another hydrophobic residue, Ile58, is exposed on the loop between Strands D and E.

Residues 41 to 62 from both a2umup and model 4 are shown in figure 4–27.

Residues 63 to 79

Residues Glu63 to Lys73 form a long β-strand (strand D) which interacts primarily with the preceeding strand C. Residues Glu63 to Tyr72 form hydrogen bonds with Glu60 to Ser51 (strand C). Residues Tyr72 and Lys73 form hydrogen bonds with Phe81 and Tyr80 (strand E). This limited interaction with strand E allows the orthogonal arrangement of the two β-sheets to occur. Strand D therefore marks the end of the first β-sheet, formed by strands A, B, C and D. Strand D is terminated by a broad bend which becomes a β-turn at Asp77 and

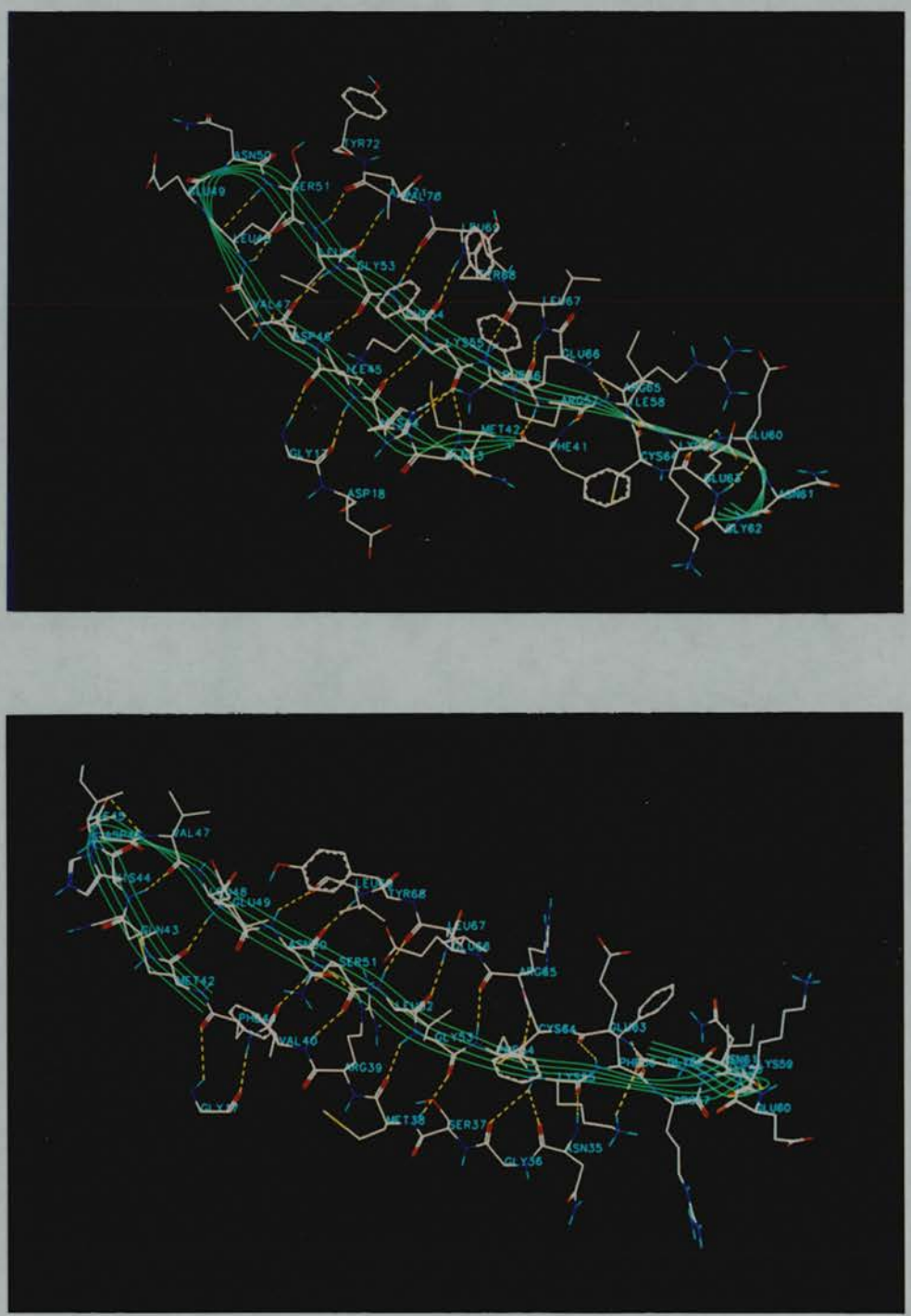


Figure 4–27: Residues 41 to 62 from a2umup (top) and model 4 (bottom).

Gly78, stabilised by hydrogen bonding between Glu76 and Glu79. The solitary disulphide bridge in a2u is formed between Cys64 and Cys157.

Residues Glu63 to Leu69 form a shorter β-strand in model 4 compared to a2umup. Hydrogen bonds are formed between Glu63 to Leu69 and Phe56 to Asn50 (strand C). There is no hydrogen bonding between the C-terminal end of strand D and the N-terminal end of strand E. Strand D is terminated, somewhat prematurely by a broad bend which becomes a β-turn at residues Asp77 and Gly78, this bend is stabilised by hydrogen bonding between Glu76 and Glu79. A sidechain hydrogen bond is formed between Asp77 OD2 and the mainchain nitrogen of Asn9. As with a2umup, strand D marks the end of the first β-sheet, formed from strands A, B, C and D. The stabilising disulphide bond between Cys64 and Cys157 is also seen.

Residues 63 to 79 from both a2umup and model 4 are shown in figure 4–28.

Residues 80 to 95

Residues Tyr80 to Glu83 form a short β-strand of length 4 (strand E). Residues Tyr80 and Phe81 form hydrogen bonds with Lys73 and Tyr72 respectively (strand D). The major interaction is with the following β-strand (strand F), with Tyr80 to Glu83 hydrogen bonding with Phe90 to Gly87. A short bend formed by residues Tyr84 to Gly86 is followed by a longer β-strand from Gly87 to Thr95. This strand interacts with the preceding strand (E) and the following strand G. Residues Gly87 to Ile92 form hydrogen bonds with Phe108 to Phe103, the ladder is broken briefly at Leu93, but continues as Lys94 to Thr95 hydrogen bond with Met102 to Val101. The loop between Strands E and F is stabilised by hydrogen bonds from Asp85 OD2 and Ser37 OG1, and from Asn88 ND2 to Tyr84 O. Residue Thr91 forms a hydrogen bond from its sidechain oxygen (OG1) to one of the ring nitrogens (ND1) of His104.

In model 4 strand E begins a residue earlier at Glu79, this short β-strand extends to Val82. Residues Glu79 to Val82 hydrogen bond with Thr91 to Asn88 (strand F). A 5 residue turn terminates strand E and then leads into strand F. This turn

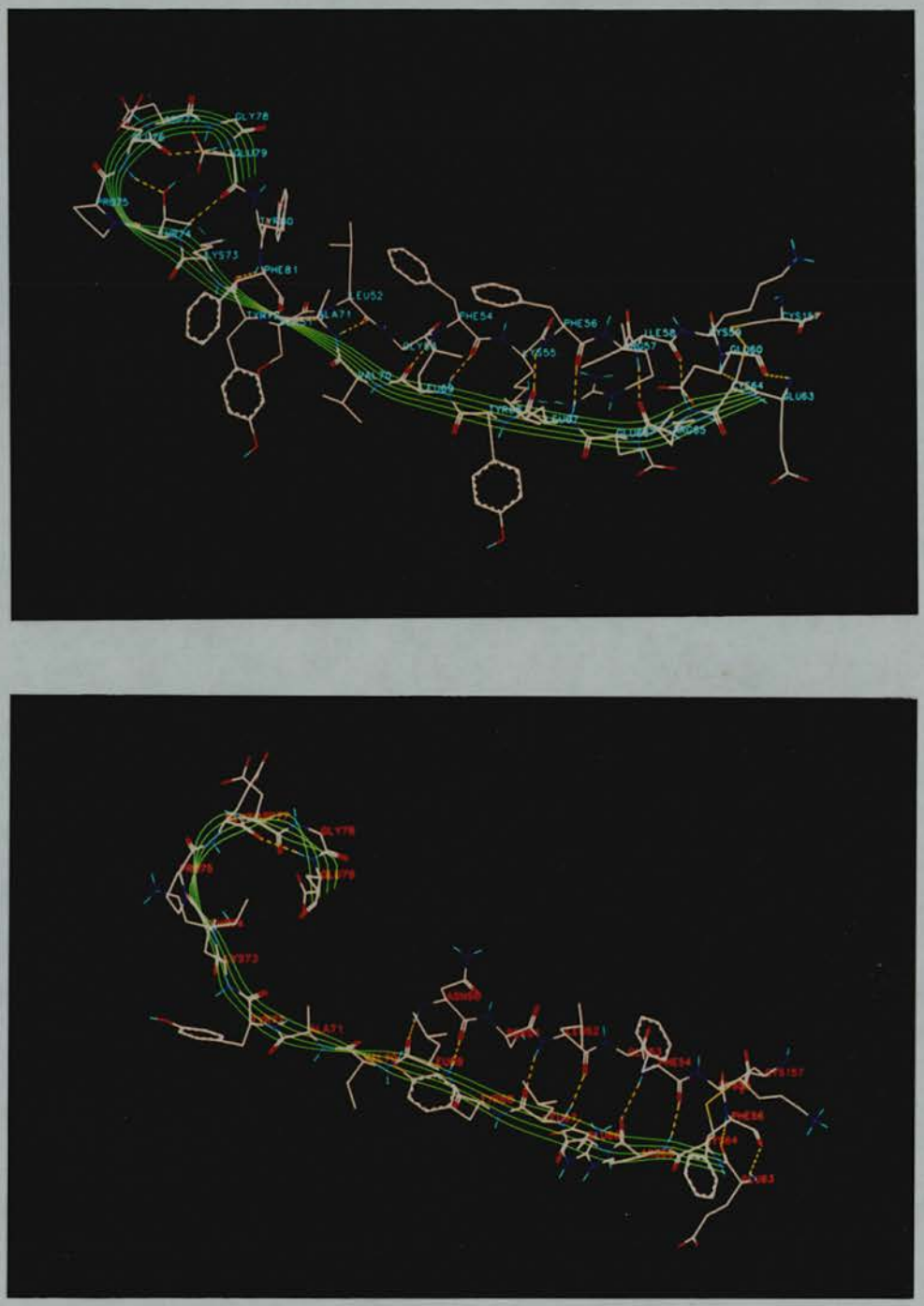


Figure 4–28: Residues 63 to 79 from a2umup (top) and model 4 (bottom).

is stabilised by hydrogen bonding from Tyr84 to Gly87, and Glu83 to Gly87. Strand F, formed by Asn88 to Thr91, is significantly shorter when compared to strand F in a2umup (4 residues as opposed to 8). Residues Thr89 to Thr91 form hydrogen bonds with 81 to 79 (strand E) and 106 to 104 (strand G). This strand is terminated by an irregular loop, which eventually leads into strand G.

Residues 80 to 95 from both a2umup and model 4 are shown in figure 4–29.

Residues 96 to 121

Residues Asp96 to Arg99 form the bend which connects strands F and G. This is also the T-D-Y motif seen in most of the lipocalycins. The loop formed by residues is stabilised by a mainchain hydrogen bond between Asp96 N and Tyr100 O. The sidechain conformation of Tyr100 is unusual - the sidechain is sandwiched between the edge of strand G and the residues just prior to strand A. Strand G is formed by residues Tyr100 to Lys109. Residues Val101 to Phe108 form hydrogen bonds with the preceding strand F. Residues Tyr100 to Lys109 form hydrogen bonds with Gly121 to Glu112 (strand H). Strand G and H are linked by a short β-turn, stabilised by a hydrogen bond between Lys109 and Glu112. Residues Glu112 to Gly121 form a long β-strand (10 residues). This strand interacts with the preceeding strand G and the first strand A. Residues Leu116 to Gly121 form hydrogen bonds with Ser26 to Phe20 (strand A). Strand H marks the end of the second β-sheet, formed from strands E, F, G and H. The interaction between strands H and A closes the two β-sheets to form a continuous β-barrel.

The area around strands G and H is less regular in model 4. This T-D-Y loop is not stabilised by any mainchain hydrogen bonding. Instead, the loop is actually interdigitated by the sidechain of Arg122. This allows the one of the terminal nitrogens (NH1) of Arg122 to hydrogen bond with Asp96 OD2. In addition, the conformation of Tyr100 is not as seen in a2umup, the sidechain is not trapped between the sheet and start of the barrel, instead it points out into the solvent. Residues Tyr97 to Arg99 therefore form a bend which leads into strand G -

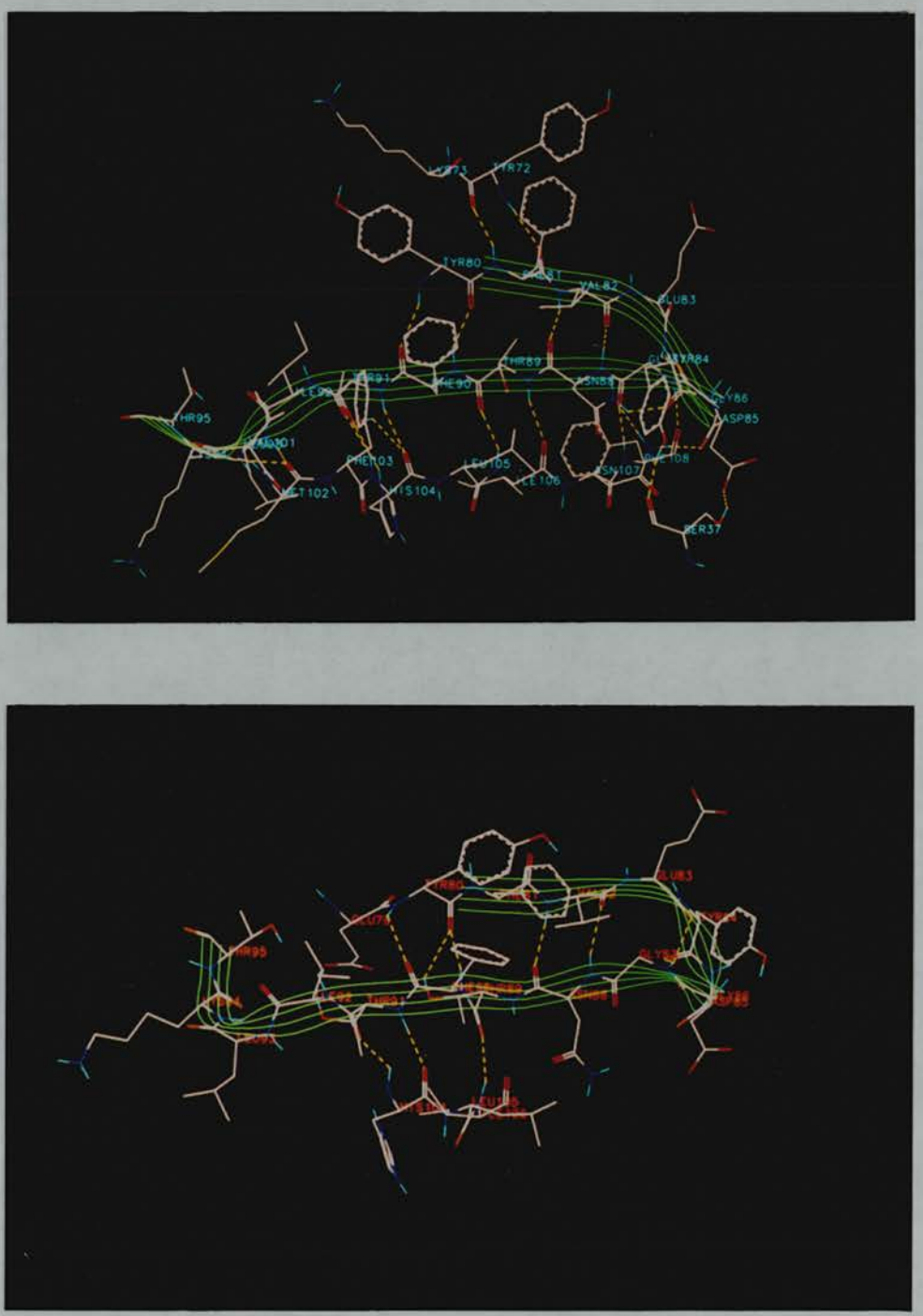


Figure 4–29: Residues 80 to 95 from a2umup (top) and model 4 (bottom).

formed by residues Val101 to Ile106. Residues His104 to Ile106 form hydrogen bonds with Thr91 to Thr89 (strand F), while Val101 to His104 hydrogen bond with Leu119 to Leu116 (strand H). This strand is terminated by a long bend (Phe108 to Gly111) and turn (Glu112 to Thr113) which lead into strand H which is shorter than in a2umup. Strand H is formed from residues Leu116 to Tyr120 and is therefore only 5 residues long. Hydrogen bonds are formed between Leu116 to Leu119 and His104 to Val101 (strand G) and Ala25 to Ser21 (strand A). As with a2umup strand H marks the end of the second β-sheet and also interacts with strand A to form a closed β-barrel topology.

Residues 96 to 121 from both a2umup and model 4 are shown in figure 4–30.

Residues 122 to 142

Residues Arg122 to Leu126 form a broad bend which leads into the solitary α-helix. The hydrophobic part of Arg122 packs against the face of Trp19. Residue Leu126 points inwards to the helix/sheet interface. The helix runs from Ser128 to Glu139 and is stabilised by the usual mainchain hydrogen bonding between residues at i and $i+4$. The α-helix is terminated by a β-turn which is stabilised by hydrogen bonding to residues within the helix, Cys138 and Glu139 interact with His 141, Gly142, and Ile143. Residues Ile130, Phe134, and Leu137 form a hydrophobic face to the inner surface of the α-helix. This face packs against the outer face of strands F, G, H, A, and I where several hydrophobic residues lie. The α-helix packs at an angle of approximately 15° against strand H. The hydrophobic regions of Lys131 and Lys133 also pack against to sheet at the helix/sheet interface.

Strand H in model 4 is terminated by a well defined β-turn at Gly121 and Arg122, stabilised by hydrogen bonding between Tyr120 and Thr123. This turn becomes a less well defined bend which leads into the α-helix. Residue Leu126 points inwards to the helix/sheet interface. The helix extends from Asp129 to Cys138 and is stabilised with the usual mainchain hydrogen bonding. The helix has a slight kink at residues Cys138 and Glu139 which forces a β-turn

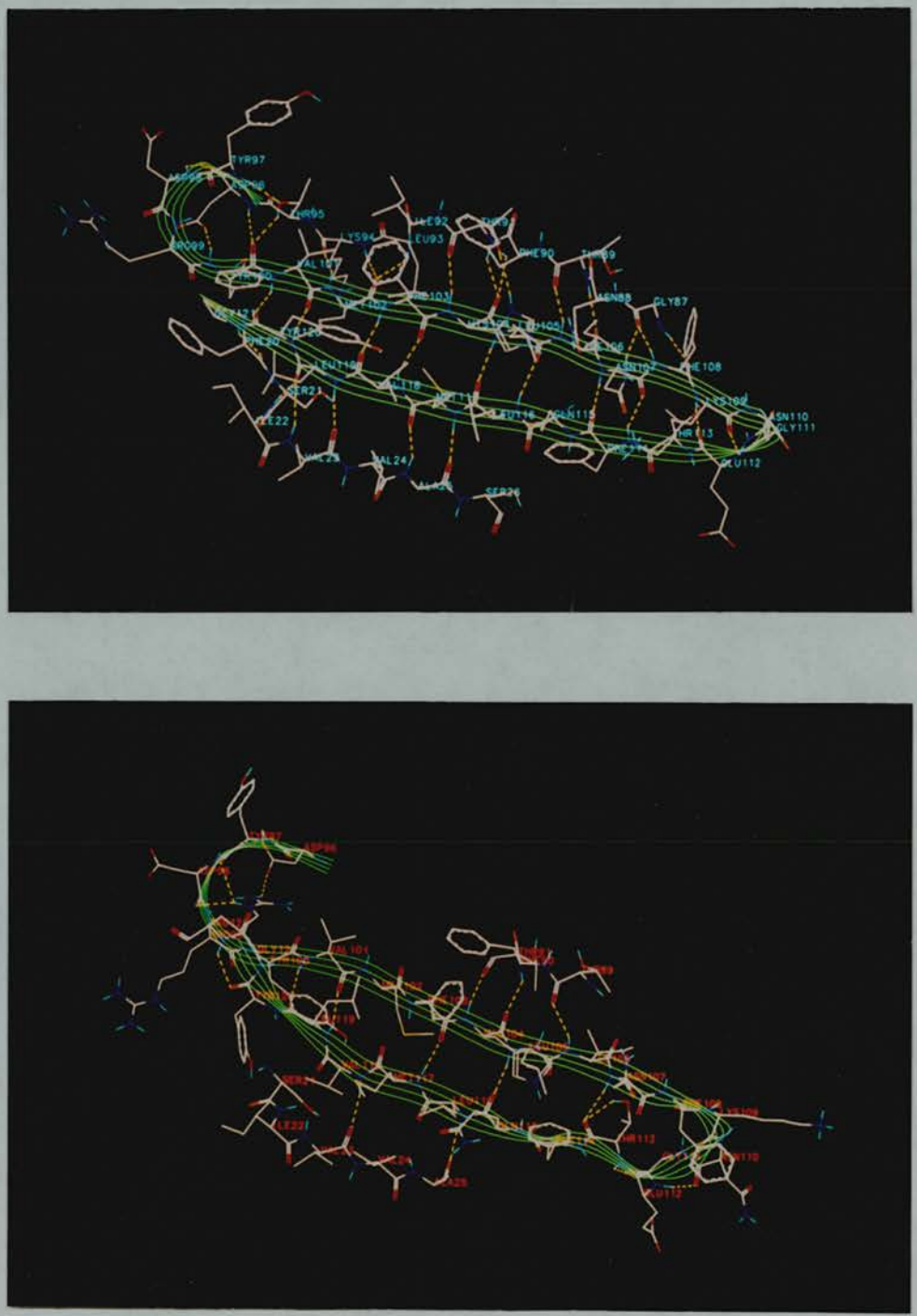


Figure 4–30: Residues 96 to 121 from a2umup (top) and model 4 (bottom).

comformation. This kink is caused by mainchain hydrogen bonding from Leu137 to Ala140, and a sidechain hydrogen bond from Lys136 NZ to Glu139 OE1. Residues Ile130, Phe134, and Leu137 form a hydrophobic face to the inner surface of the helix. However, the helix is shifted such that it lies over strand G (a 5 Å lateral shift away from strand H). The helix is also shifted 2.5 Å towards the N-terminus. The helix therefore packs against strands F, G, H, and A, at an angle of approximately 15°. The hydrophobic part of Lys133 packs at the helix/sheet interface, but Lys131 is shifted such that its sidechain is exposed to the solvent.

Residues 122 to 142 from both a2umup and model 4 are shown in figure 4–31.

Residues 143 to 157

The β-turn formed at the termination of the α-helix leads into a short stretch of 3_{10}-helix comprising residues Arg145 to Asn147. This 3_{10}-helix allows the mainchain direction to change such that it runs anti-parallel to the α-helix. The 3_{10}-helix leads directly into a short β-strand formed by residues Ile148 to Asp150 (strand I). This strand hydrogen bonds with residues Ala25 to Val23 (strand A). The mainchain trace continues from strand I to the C-terminus, with residues Thr152 to Asp155 forming a repeated β-turn, which has the appearance of an *alpha*-helix. This turn is stabilised by hydrogen bonding from Leu151 to both Thr154 and Asp155, and from Lys153 to Arg156. Sidechain hydrogen bonds are seen from Asp155 OD1 to Thr152 O, and from Arg156 NH2 to Gly62 O. The crystallographic structure of MUP is terminated at Cys157, which forms a disulphide bond with Cys64. It is assumed residues between Cys157 and the C-terminus are too disordered to be observed.

Residues Gly142 to Arg145 form a loop which leads in the final β-strand. Residue Ile143 lies exposed to solvent on this loop. Residues Asp146 to Ile148 form a β-strand (strand I), which forms hydrogen bonds to Ala25 to Val23 (strand A). Residue Ile148 is exposed to the solvent in this position. The mainchain then continues towards the C-terminus through a repeated β-turn.

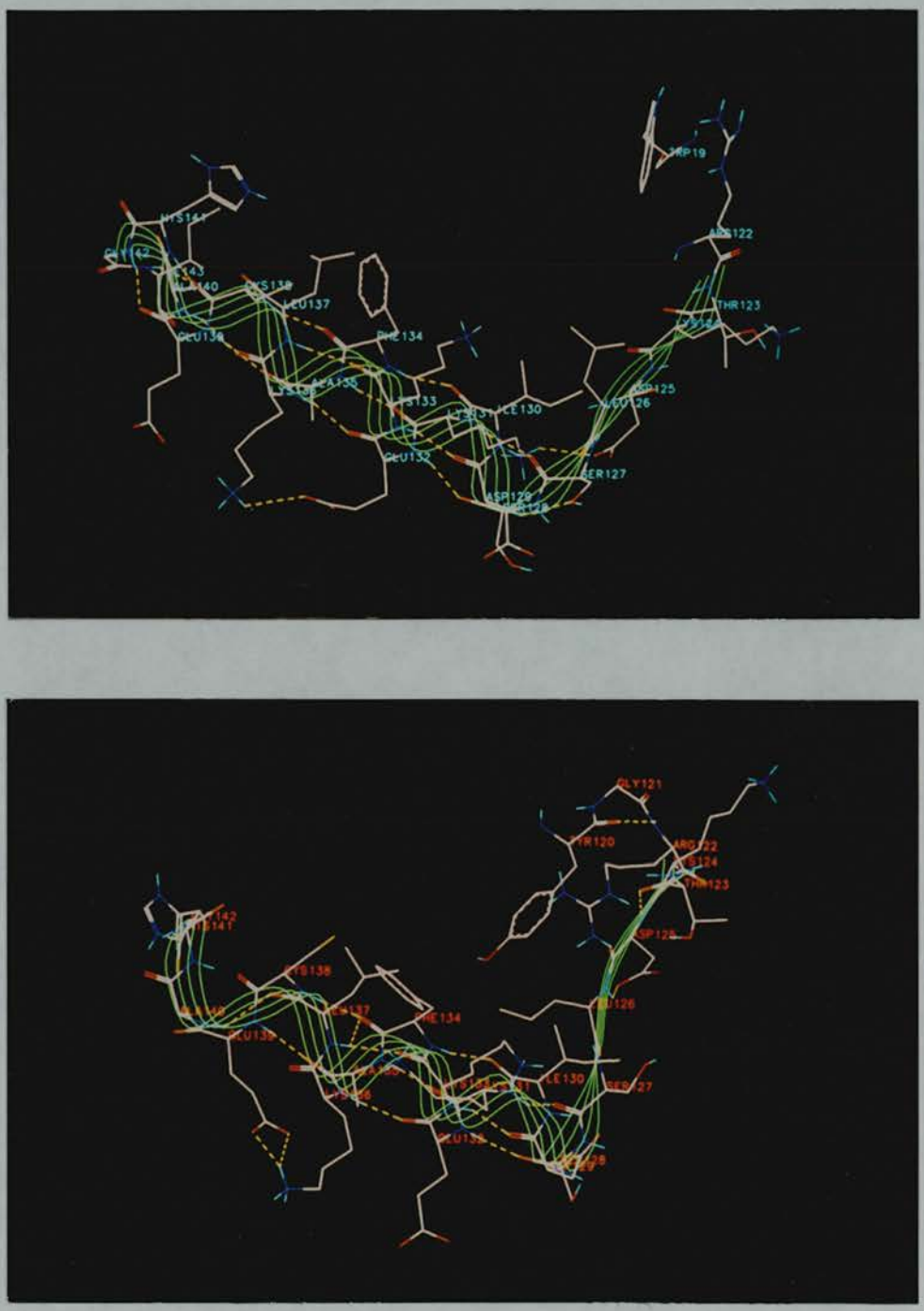


Figure 4–31: Residues 122 to 142 from a2umup (top) and model 4 (bottom).

This turn is stabilised by hydrogen bonding from Leu151 to Asp155, and from Lys153 to both Arg156 and Cys157. Sidechain hydrogen bonding is observed between Asp155 OD1 and Leu151 O. The conformation of the turn is close to that of a α-helix. The model is terminated at Cys157 which forms a disulphide bond with Cys64.

Residues 143 to 157 from both a2umup and model 4 are shown in figure 4–32.

4.5 Structural analysis of the Lipocalycins

The analysis of models of a2u based on both MUP and other lipocalycins prompted the analysis of all the reliable lipocalycin structures available at the time. The aim was to determine which specific features are common to all these structures. Particular attention was given to those non-bonded interactions which stabilised the structures, and any consistant deviations from normal protein geometry.

The refined structures for several lipocalycins were analysed: RBP (RBP.dat), BBP (BBP.dat), INSEC, and MUP. The coordinates RBP.dat and BBP.dat were obtained from the PDB. RBP.dat is the final structure of RBP refined at 2.0 Å resolution with a final R-factor of 18.1% for data between 8.0 and 2.0 Å. BBP.dat is the final structure of BBP refined at 2.0 Å resolution with a final R-factor of 20%. The structures were superimposed using lsq_explicit then lsq_improve in the program O (table 4–13). The α-carbon coordinates of the superimposed structures were analysed to determine conserved positions between the proteins. This was achieved by modification of the α-carbon averaging program described earlier. The program was modified to output the identity of those residues whose α-carbons were within a sphere of a user specified radius for all input proteins. If there were n input structures and a template structure (the first input structure) of m residues the output listing has several groups of n residues, at most there are m groups. A two pass selection scheme was used: firstly the closest α-carbons, from all other structures, to a target α-carbon atom

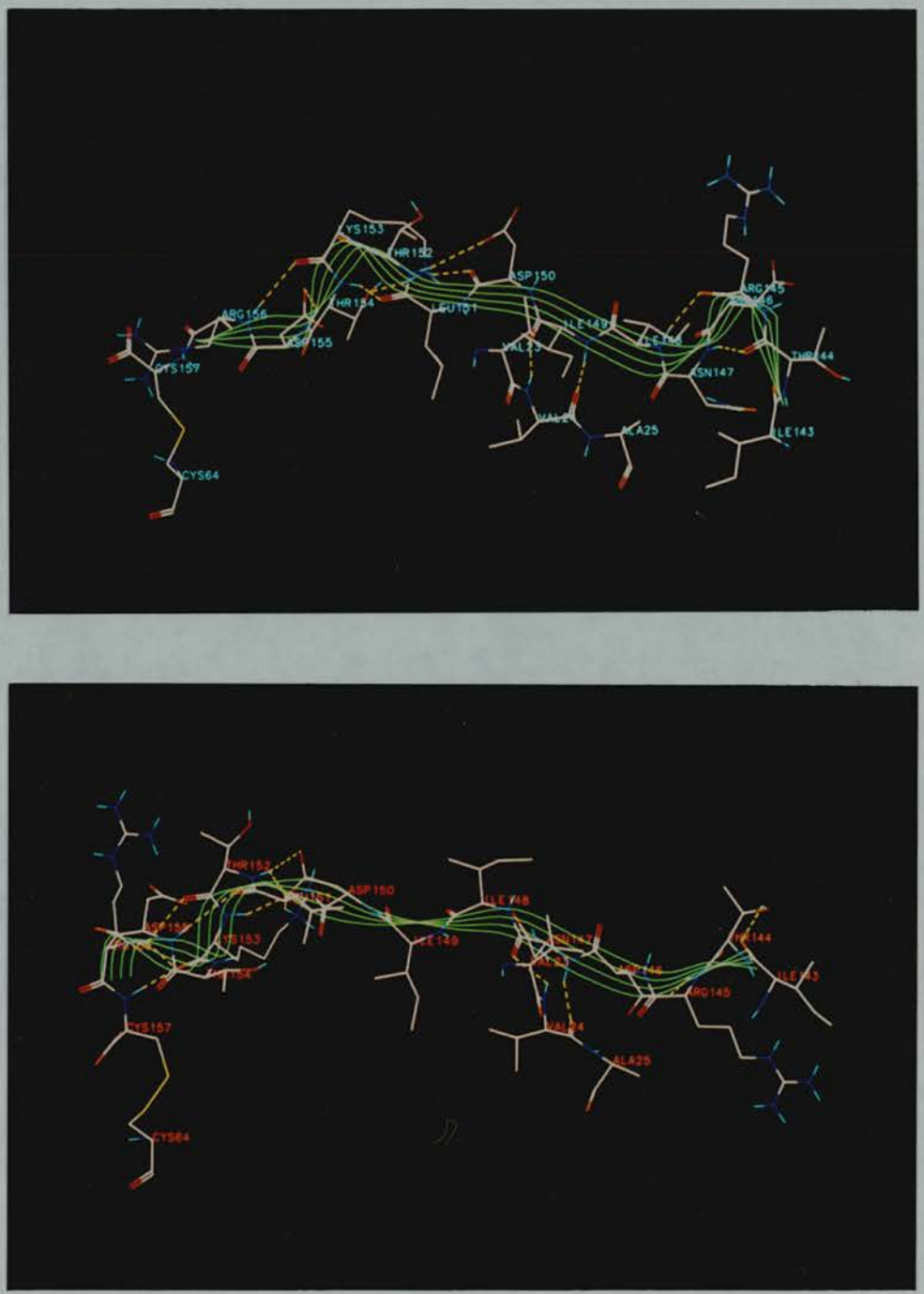


Figure 4–32: Residues 143 to 157 from a2umup (top) and model 4 (bottom).

	MUP	BBP	INSEC
RBP	1.799 (100)	1.731 (99)	1.902 (108)
MUP		1.854 (90)	1.837 (83)
BBP			1.557 (140)

Table 4–13: Analysis of superposition of RBP, MUP, BBP, and INSEC. Figures are rms deviations in Å for those α-carbon atoms matched by the lsq_improve algorithm (the number in brackets).

in the template molecule are selected. If this process selected an α-carbon from all other structures the centre of geometry for those atoms is calculated. Secondly, if the selected α-carbon atoms for this template residue were all within the user specified cutoff distance of the centre of geometry for those atoms the α-carbons were selected. The atom number and residue type for each structure were written along with the rms deviation of the α-carbons from the centre of geometry. In addition the coordinates of the selected centres of geometries were written for visual analysis. This procedure was applied to the superimposed structures of RBP, BBP, MUP and INSEC using a cutoff radius of 2.0 Å. The structurally conserved residues were then used to align the amino acid sequences of RBP, BBP, MUP and INSEC (figure 4–33).

The core of structurally conserved α-carbon atoms within a cutoff of 1.25 Å predominantly consisted of the second β-sheet (figure 4–34). Alpha-carbon atoms from strands G, H, A and I were conserved even at 1.0 Å cutoff. As the cutoff was relaxed to 1.5 Å elements of the first sheet were seen, mainly strands B and C. It is interesting to note that the conserved sequence motifs (G-x-W and T-D-Y) did not appear as being structurally conserved at 1.0 or 1.25 Å cutoffs. The G-x-W motif is only seen at a cutoff of 1.5 Å. These motifs in the four structures superimposed very closely when aligned in isolation (¡0.5 Å). The placement of these regions with respect to the rest of the molecule varies for each structure. Therefore when the structures are compared as a whole these motifs do not superimpose as well.

Detailed analysis of the structures indicated certain residues which occur consistantly, these are in agreement with those discussed in the analysis of the RBP structure (Cowan *et al.*, 1990).

Figure 4–33: Sequence alignment of RBP, BBP, INSEC, and MUP incorporating information from three dimensional alignment of the structures.

Figure 4–34: Core α-carbon atoms computed after superpostion of RBP, BBP, MUP and INSEC.

G-x-W motif

The glycine residue is highly favoured at this position in the lipocalycin sequences, only rat A1GP lacks this glycine (Schmid, 1975). The glycine adopts a preferred conformation, with ϕ/ψ angles of approximately 120°, -160°.

Tryptophan is absolutely conserved in this motif across all lipocalycin sequences. This residue forms mainchain hydrogen bonds to Ala43 (RBP), Ala46 (BBP), Ala46 (INSEC) and Glu43 (MUP). These residues are part of strand B in each structure respectively, and lie within 1.5 Å of one another in the structural superpostion. The sidechain atoms of the tryptophan residue form part of a hydrophobic cluster. Trp24 in RBP packs with Phe20, Tyr114, Ala115, Phe137, and Arg139 (packing across the face of the tryptophan ring). Trp19 in MUP packs with Ile15, Leu42, Ile45, Leu101, and Arg122 (packing across the face of the ring). Trp27 in BBP packs with Tyr23, Tyr48, Ile114, Leu135, and Arg137 (packing across the face of the ring). Trp27 in INSEC packs with Phe23, Tyr48, Ala114, Leu135, and Lys137 (packing across the face of the ring). The formation of the hydrophobic cluster centered around the tryptophan sidechain would seem essential to the lipocalycin topology. The stacking of the hydrophobic portion of

a long sidechain, such as arginine or lysine, across the face of this tryptophan is well conserved. Other sequences have either an arginine or lysine in the equivalent position. This interaction between a residue at the end of strand H (arginine or lysine) with a residue at the start of strand A (tryptophan) may be vital in stabilising the β-barrel topology.

The tryptophan residue is generally followed by a bulky amino acid, often with a ring-based sidechain. The residue immediately following the tryptophan forms mainchain hydrogen bonds with Ser138 (RBP), Gly121 (MUP), Ser136 (BBP), and Ser136 (INSEC). This again strengthens the interaction between strand H and strand A. Tyr25 in RBP packs against the face of Pro141. His20 in MUP similarly packs against the face of Pro124. Trp28 in BBP is sandwiched between the hydrophobic parts of Lys26 and Lys139. His28 in INSEC is loosely sandwiched between Ile45, Lys139, and the face of Phe170. In all cases there is an interaction with the residue two places C-terminal to the well conserved arginine or lysine which packs with the tryptophan. In general this residue is a proline, lysine, arginine or hydrophobic. Again this suggests that this packing interaction is important in stabilising the topology.

T-D-Y motif

This sequence is well conserved amongst the lipocalycin sequences, being notably absent in most of the lipocalycins involved in olfaction. Thr109 in RBP forms hydrogen bonds from OG1 to the mainchain oxygen of Tyr114 and mainchain nitrogen of Tyr111. Thr95 in MUP forms hydrogen bonds from OG1 to the mainchain oxygen of Phe100, and the mainchain nitrogen of Tyr97. Thr108 in BBP forms hydrogen bonds from OG1 to the mainchain oxygen of Tyr113, and the mainchain nitrogen of Tyr110. Thr108 in INSEC forms hydrogen bonds from OG1 to the mainchain oxygen of Tyr113, and the mainchain nitrogen of Tyr110.

Asp110 in RBP forms hydrogen bonds OD1 to the mainchain nitrogens of Thr113 and Tyr114. Asp96 in MUP forms hydrogen bonds from OD2 to the mainchain nitrogens of Asn99 and Phe100. Asp109 in BBP forms hydrogen bonds from

OD1 to the mainchain nitrogens of Asn112 and Tyr113. Asp109 in INSEC forms hydrogen bonds from OD1 to the mainchain nitrogens of Asn112 and Tyr113.

The sidechain of the tyrosine residue, or its equivalent, in all four structures is in an unusual conformation, being wedged between the second sheet and the residues at the start of the barrel (prior to the G-x-W motif). This distorts the backbone conformation of the loop, allowing hydrogen bonding between the mainchain nitrogen and OG1 of the preceding threonine.

The two residues immediately following the tyrosine of the T-D-Y motif are diverse in nature, and form no major mainchain or sidechain interactions. However, the third residue after the tyrosine is strongly favoured to be a tyrosine or phenylalanine. In all four structures the hydrophobic ring points back across the face of the loop, thus lying at the interface between the α-helix and the second β-sheet.

Other Residues

A leucine residue is strongly favoured on the bend which links strand H and the α-helix. Leu144 in RBP, Leu126 in MUP, Leu141 in BBP, and Leu141 in INSEC, all point from the bend into the helix/sheet interface. Presumably this residue is important in either stabilising the interaction between helix and sheet or alternatively in promoting the formation of an α-helix at this point.

4.6 Discussion

The modelling of a2u produced one model which appeared to be a candidate for the native structure. That model 4 was the best model was inferred from comparison with other models produced using different modelling methods. Analysis of the secondary structure of this model indicated that it possessed the overall topology of the lipocalycin family. However, detailed comparison of the model with the structure of MUP indicated that there were major differences. In

addition, analysis of the lipocalycin structures determined to date indicated that any lipocalycin model should have certain structural features. The best model of a2u produced did not show many of these and therefore could not be considered to be a native-like structure for a2u. The presence of charged residues in the hydrophobic core of the model also suggested that the model was incorrect (Blundell *et al.*, 1987).

In particular the model had the sidechain conformation of the absolutely conserved tryptophan (Trp19) rotated 180° away from the expected position. Also, the mainchain and sidechain conformation of the highly conserved T-D-Y loop are not like other lipocalycin structures. This reason for these differences must lie with the placement of sidechains during model building. The model building procedure (method 4) started from conserved C-α coordinates, from which mainchain atoms were built. Subsequently sidechains were added automatically using SYBYL. This uses the position of the mainchain atoms to determine the orientation of the sidechain. At this point any local disturbances to the mainchain conformation will result in gross changes to the positions of the sidechains. Subsequent energy minimisation of the model cannot hope to overcome these problems. The local minima found by the minimisation procedure will be truely local - rotation of sidechains by 180° through stable secondary structure elements is not possible. Therefore, the conformation of these conserved residues should have been fixed, by including the whole residue from the start of the modelling. The strong sequence/structure conservation of a few residues provides a few fixed points around which the rest of the model can be built.

The attempt to improve model 1 by use of a simulated annealing procedure also did not result in the native structure. It is interesting to note that some elements of the model did improve. The α-helix became better defined (as assessed by the program DSSP) as did some of the tighter β-turns. However, other features became less well defined - most β-strands became distorted. These observations can be interpreted in several ways. It could be argued that the simulated annealing technique is not suitable for the refinement of protein models.

However, the use of simulated annealing in protein modelling has been reported elsewhere and appears to have be successful (Nilges and Brünger, 1991). It is possible that SA is only successful when the starting model is close to the correct, native, protein structure. Intuitively this seems reasonable, because incorrect regions of the model will move over the time of the simulation to try and overcome bad local contacts. This movement is more likely to distort the structure than take it towards the native structure. As with energy minimisation, the radius of convergence is still too small for SA to overcome gross inaccuracies in the model (for example the misthreading of strands).

Chapter 5

Parallel Processing

5.1 Background

There are many scientific problems which require a prohibitively large amount of computation time. As a consequence faster computers are always being developed, usually resulting in a lower cost per performance unit. Computing appears to be at a crossroads in terms of future development. Will parallel processing replace the more conventional serial processing in the future?

5.1.1 Computer Architecture

Flynn's taxonomy of computer architecture is based upon the way a machine relates its instructions to the data being processed (Hockney and Jessop, 1981). This is perhaps a simplistic classification when machines are considered in detail but it uses only four groups, which highlight the conceptual differences between machines. A stream is defined here as a sequence of items, instructions or data, executed or operated on by a processor.

- **S**ingle **I**nstruction stream/**S**ingle **D**ata stream - **SISD**. This is the conventional, serial, von Neumann computer. One stream of instructions is executed by one processing unit which acts on one stream of data. As a consequence the processing unit only considers one instruction and one data item at a time.

- **S**ingle **I**nstruction stream/**M**ultiple **D**ata stream - **SIMD**. This machine has a single stream of instructions which are applied to many data items at the same time. This implies that there is more than one processing unit in such a machine, the maximum number of data items that can be processed at one time is the same as the number of processors.

- **M**ultiple **I**nstruction stream/**S**ingle **D**ata stream - **MISD**. In this machine many instructions act on a single data stream at the same time.

- **M**ultiple **I**nstruction stream/**M**ultiple **D**ata stream - **MIMD**. Multiple instruction streams imply the existence of several processing units. Each processing unit acts on its own data stream. Processing units may process an individual data stream exclusively or data streams may be swapped by communication of data items between processors.

These classifications are very broad, and do not easily accommodate the major pipelined vector supercomputers such as CRAY-XMP, Convex C220, Fujitsu VP and Cyber 205. However, for the parallel computers considered here the classification is appropriate. Firstly, the Meiko Computing Surface (CS) is presented as an example of a MIMD machine. The Connection Machine (CM-200), manufactured by Thinking Machines Corporation, is presented as an example of a SIMD machine.

5.1.2 Parallel Concepts

The first valve computers carried out all calculations serially, even the bit addition between two bytes. Since then the use of integrated circuit technology has allowed such operations to occur on all bits in parallel. In addition calculations and input/output are overlapped. Therefore, all computers are parallel to some degree although this is not apparent to the programmer or user. At present we must consider parallel computers as those whose parallel nature is evident to the programmer, altering the way in which programs must be written. Obviously the parallel nature of a particular machine depends on the hardware

architecture. Some architectures are fixed, whereas others can be changed dynamically. Memory may be specific to a processing element or may be global. As the number of processing elements increases it is difficult to implement global memory in hardware, instead local memory can be made to appear global by software. Increasingly the architecture of machines is less important to the programmer as high level languages automatically deal with dynamic changes and memory distribution.

5.1.3 Meiko Computing Surface

This machine consists of many independent RISC processors (the Inmos transputer; INMOS Limited, 1988) each with its own local memory and a means to communicate with other transputers across reconfigurable electronic links (figure 5–1). A transputer has four 20 Mbit/s bidirectional communication links which places limits on the topology of connectivity between processors. Each transputer has its own unique code which is loaded at run time from a host processor. There is no shared, global memory; only local memory, 4 Kbytes of which are on chip. Each T800 transputer has a peak performance rating of 1.5 MFLOPS, but in practice 0.8 MFLOPS sustained performance is achieved (Raine *et al.*, 1989). The largest domain available on the Edinburgh Meiko Computing Surface (ECS) is 131, and therefore offers a maximum performance of 105 MFLOPS.

As a consequence of the independent nature of each processor, communication must be by explicit passing of data items - message passing. There is no hardware synchronisation of processor calculations, therefore the machine is truely concurrent. Synchronisation of calculation is possible in software by dependence upon data communication. The model for programming is communicating sequential processes (Hoare, 1985). A process may be a whole transputer or there may be many virtual processes running on one transputer. Obviously maximum parallelism only occurs when each process runs on an individual processor. The concurrent sequential process model can be

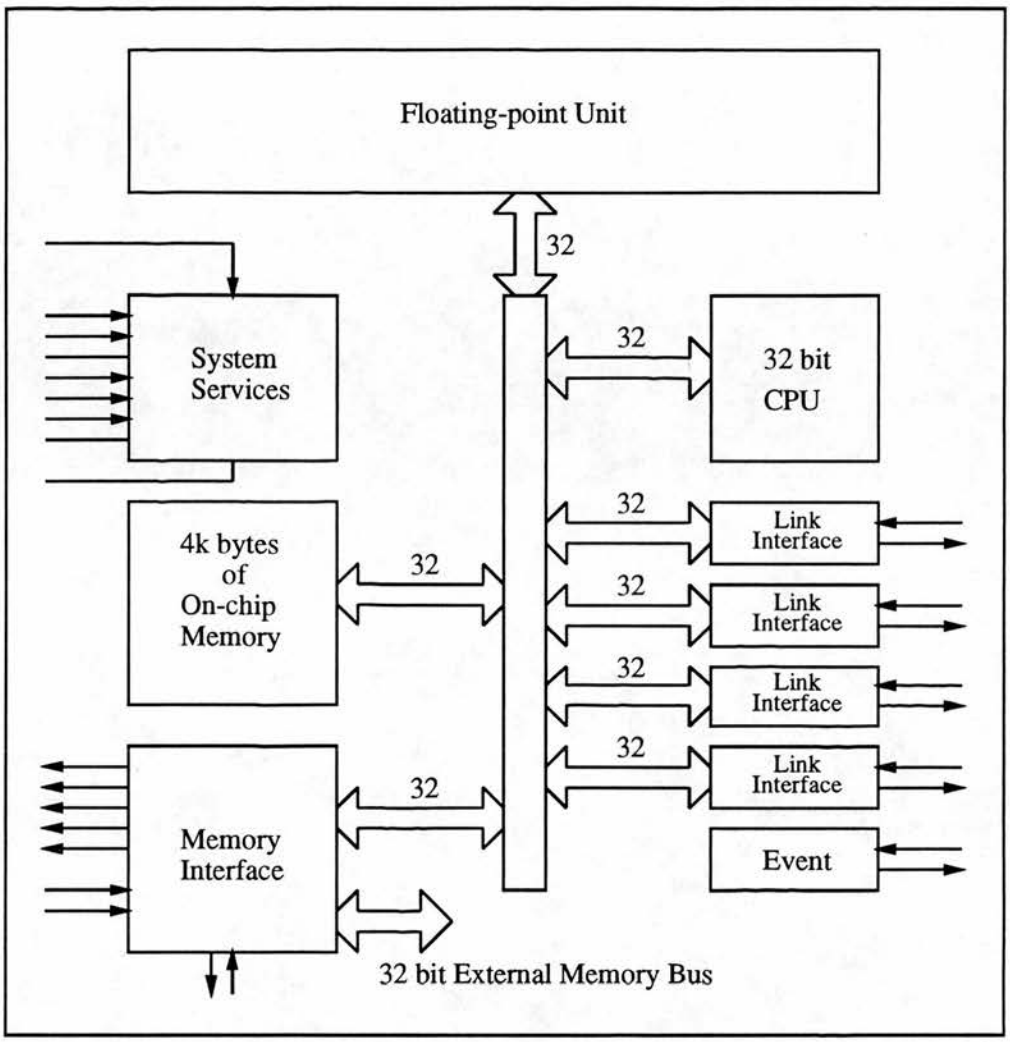


Figure 5–1: The Inmos T800 Transputer.

implemented within standard languages such as Fortran-77 and C by the use of library calls. Alternatively, parallel languages such as Occam have been developed alongside the hardware. This language has message passing and parallel constructs as part of its high level instruction set (Jones and Goldsmith, 1988). However, hardware developments and the nature of many of the programs means that standard languages such as Fortran-77 and C are favoured.

5.1.4 The Connection Machine (CM-200)

The Connection Machine CM-200 is a data parallel computing system (Thinking Machines Corporation, 1991a). Data parallel computing associates one processor with each data element. A CM-200 parallel processing unit may contain 2, 4, 8, 16, 32, or 64K data processors, where K stands for 1024 (2^{10}). These processors are used whenever an operation can be performed simultaneously on many data objects. Data objects remain in the CM-200 memory during program execution and are operated on in parallel. A front-end computer is associated with a CM-200 system. This front-end is where code is executed; instructions are only passed to the CM-200 when parallel operations on data in CM-200 memory are to be carried out. Communication between the front-end and CM-200 is by a fast data bus.

The CM-200 implements data parallel programming constructs directly in hardware and microcode. Parallel data structures are spread across the data processing elements, with a single element stored in each processor's memory. When data parallel structures contain more data elements than the system has processors (the normal situation), the system operates in virtual processor mode. Each processor is effectively divided into several smaller processors. As the program issues parallel instructions, microcode causes it to be executed many times, once for each virtual processor. Therefore the same program can be run on different sizes of CM-200 without any changes and as the number of processing elements increases, the program runs faster.

The data processing elements are controlled by a sequencer (figure 5–2). The task of this sequencer is to decode commands from the front-end and broadcast them to the data processors, which can then execute the same instructions simultaneously. The PEs are grouped together into data processing nodes. Each node contains 32 PEs, with associated memory, an optional floating-point accelerator, and the communications interfaces for interprocessor communication (figure 5–3). Each PE houses an arithmetic-logic unit (ALU) which enables each PE to act as a bit-serial processing element. However, many scientific

applications require floating-point calculations. Therefore, a Weitek floating-point accelerator (FPA) can be associated with every 32 PEs. These FPAs are vector processors with a vector-length of 4, into which commands and data can be pipelined. Each FPA has a peak speed of 16 MFlops, therefore a 64K CM-200 has a theoretical peak speed of 32 GFlops. The two modes of using the resources of the CM-200, either using the PEs or the FPAs for computation, are distinct. The former is called the paris (**par**allel **i**nstruction **s**et) mode, whereas the latter is the slicewise mode. The term slicewise is used because data is distributed differently in memory; one bit from a 32-bit word is held per PE memory section rather than all 32-bits. In this way words are distributed in a slicewise manner across PEs. In general the optimum performance from the CM-200 can only be obtained using the slicewise execution model but there are some cases, if only bit operations are carried out, where optimum performance is obtained under the paris model of execution. The searching of DNA databases is such a case (Shane Sturrock, personal communication).

Interprocessor communication is implemented by special-purpose hardware. Message passing happens in parallel; all processors can simultaneously send data into the local memories of other processors, or fetch data from the local memories of other processors into their own. The transfer of data can be combined with arithmetic or logical operations such that the destination receives the combined result. The most general of the CM-200's communication mechanisms is the router, which allows any processor to communicate with any other processor. The topology of the router network is a boolean n-cube. For a fully configured CM-200 with 64K PEs, the network is a 12-cube. Each set of 16 PEs has one associated router node. In a 12-cube there are 4096 router nodes, each one connected to 12 other router nodes. The router nodes automatically determine the optimum path for a data item to be passed between two processors. Communications between processors that are nearest neighbours within a Cartesian grid are much more efficient than general router communication because they exploit special features of the underlying hardware. Data laid out in a regular Cartesian grid: a NEWS grid (because each PE has a

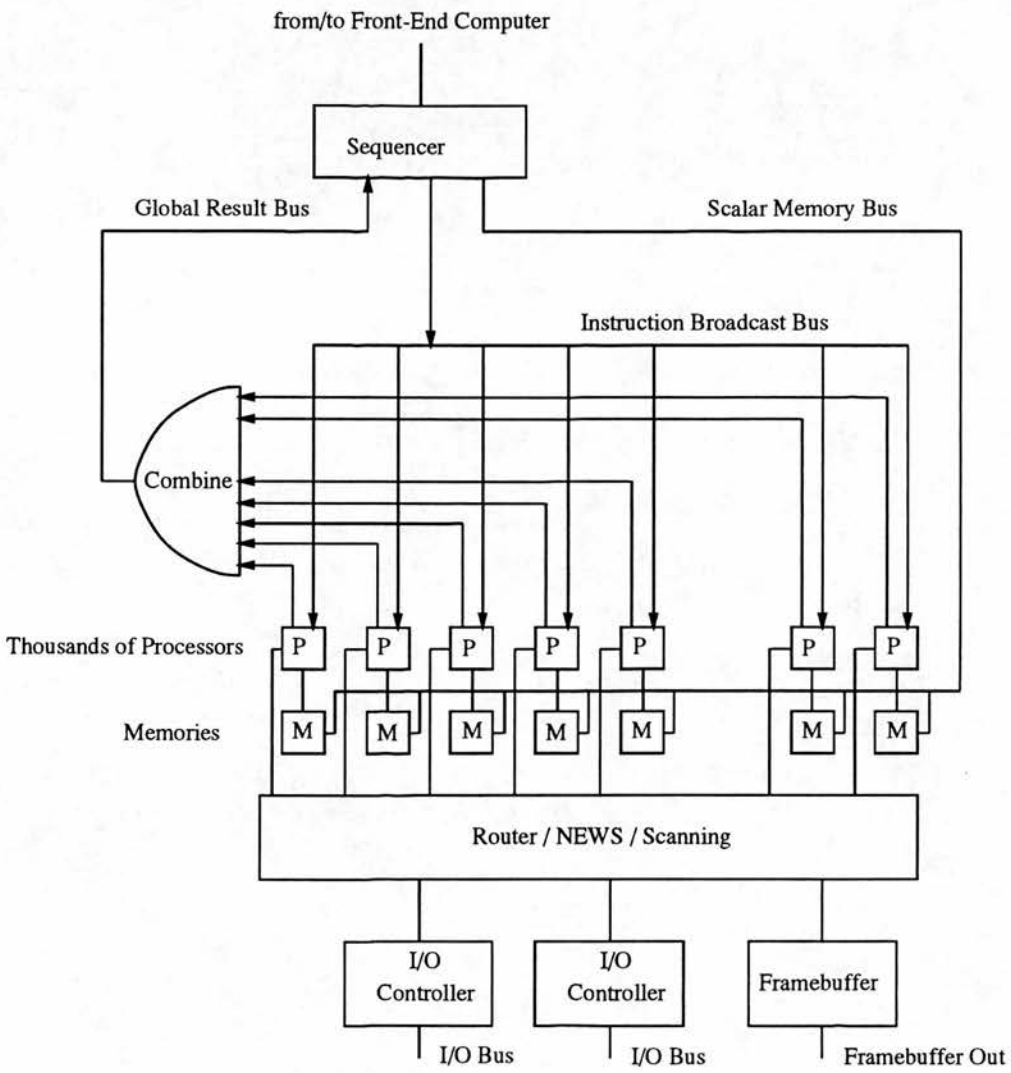


Figure 5–2: Architecture of the CM-200 Parallel Processing Unit.

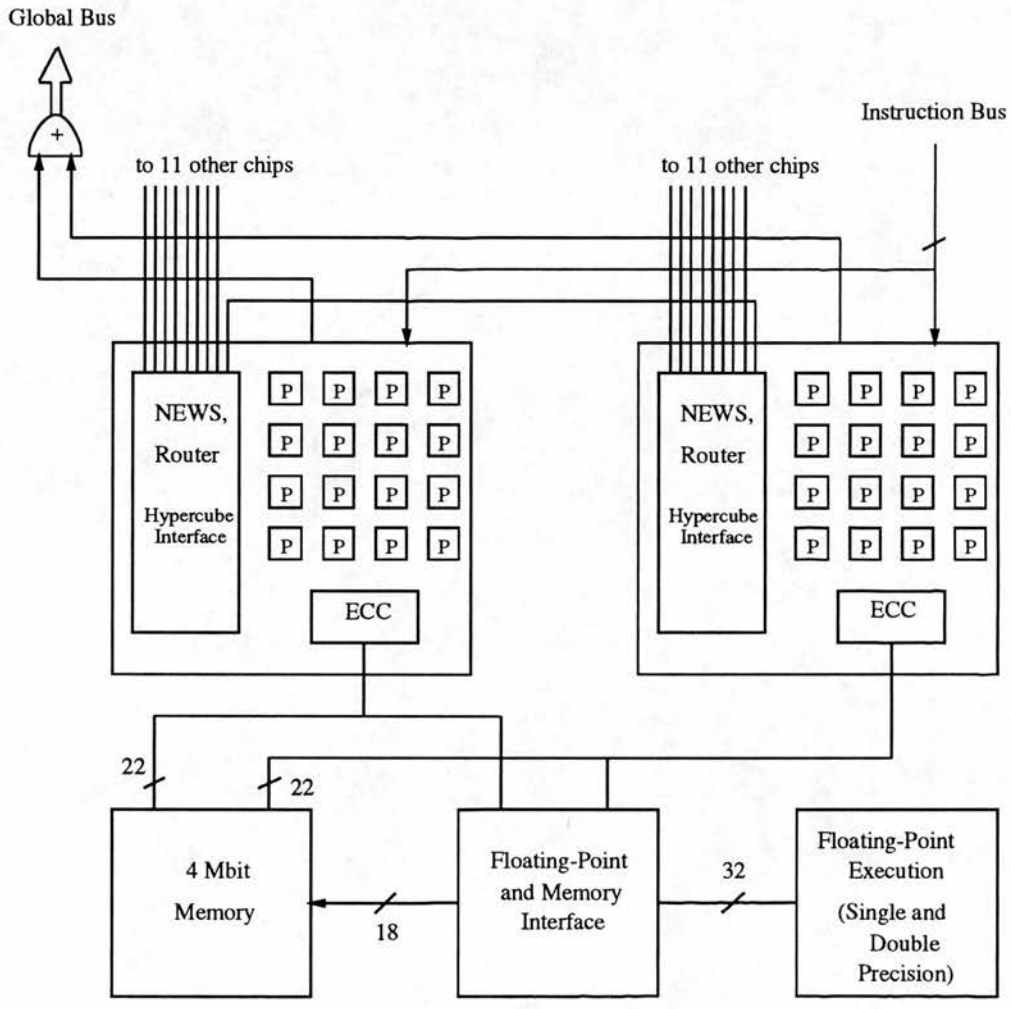


Figure 5–3: Architecture of the CM-200 Floating Point Accelerator.

north, east, west, and south neighbour), can be arranged across the PEs such that communication between neighbouring data items is at most over one hypercube wire. The data layout is so well defined that no computation of destination address is required before communication. In addition, data are arranged such that neighbouring data items actually reside on the same PE, thus obviating the need for any communication.

The CM-200 provides the hardware to carry out parallel operations on large numbers of data items, typically arrays of more than one dimension. The programming languages that can be used to exploit this hardware include CM Fortran and C*. CM Fortran is an implementation of a limited subset of Fortran 90 (Metcalf and Reid, 1990), in particular the parallel array definitions and operations (Thinking Machines Corporation, 1991b). C* is ANSI C with parallel language extensions. The machine used for the work presented later in this chapter had 16K bit-serial processors, 512 FPAs and a total of 0.5 Gbytes of memory.

The AMT Distributed Array Processor (DAP) is a SIMD parallel machine like the CM-200, and was used in some of the work presented later (section 5.2.2). It architecture is different to that of the CM-200, processors being connected in a two-dimensional grid topology. Its construction is not discussed in detail here but the underlying data parallel concepts are implemented as on the CM-200.

5.1.5 Parallel Algorithms

In order to use a particular parallel architecture efficiently, it is necessary to take account of:

- how data are distributed in memory
- how computations are distributed among processors
- communications between processors

For SIMD machines like the CM-200 the programmer has less control over these factors due to the fixed nature of the architecture. In addition, much of the communication between processors is taken care of automatically by the low level software. In dynamically reconfigurable MIMD machines, such as the CS, the programmer has control over the way processors are linked. This has the disadvantage that communication has to be explicitly described by the programmer at the high language level. There are three classes which can be used to describe the parallelisation of a problem:

- Event Parallelism
- Geometric or Data Parallelism
- Algorithmic Parallelism

These classes are considered below with particular respect to the CS and CM-200. All three classes lend themselves to dynamic, transputer based, MIMD systems, whereas only the second can be realistically used on SIMD machines such as the CM-200.

5.1.6 Event Parallelism

Each processor executes a program in isolation from the other processors. In effect, each processor is used as a separate computer, in this sense a VAX cluster uses event parallelism. There are many examples of this kind of parallelism including ray tracing (Dettmer, 1986) and calculation of the Mandelbrot set (Mandelbrot, 1982). These examples lend themselves to such parallelism because of the nature of the algorithms. In each case a portion of the screen is given to a processor for calculation. Each processor receives a copy of all of the data required to carry out the calculation on one processor. An alternative approach is the use of a task farm where several copies of the same program process their own data independently of each other. This kind of parallelism is usually very efficient provided the load placed on each processor is nearly equal.

5.1.7 Geometric/Data Parallelism

Commonly SIMD machines are used to carry out operations on all elements of an array simultaneously without communication between processors. This is true data parallel processing, for which the Fortran-90 code takes the form:

```
real, array(1000,1000) :: A, B, C

A = 1.5
B = 60.0

C = A + B
```

This data parallel processing can be implemented on both SIMD and MIMD machines. The underlying architecture of the next generation Connection Machine, the CM-5, is MIMD, however it can be programmed using both data parallel and message passing methods.

Geometric parallelism configures the processors in such a way that their geometry reflects the problem being studied. The data are decomposed onto the processors such that data items which are close together reside on processors which are close together. The problem possesses geometric parallelism if, in addition, the calculations involved are only between local data items. A classic system which lends itself easily to geometric parallelism is that of cellular automata (Wolfram, 1986). In a two-dimensional array of cells, the state of a cell depends on the states of the 4 or 8 surrounding cells. Data can be passed rapidly between processors provided there are sufficient links for communication. The CS can be electronically configured so that processors are connected in a two-dimensional grid topology. The CM-200 processors are permanently connected in a hypercube topology. In the case of a 16K processor machine this forms a 9-cube. Provided data elements of a two-dimensional array are distributed correctly across these processors, rapid communication between neighbouring elements is possible. In effect the hypercube appears to be two

dimensional to the programmer. In both cases communication between neighbouring elements is rapid, allowing the state of each cell to be updated simultaneously. Obviously, communication between processors takes some finite time, therefore not all of the time is spent calculating.

5.1.8 Algorithmic Parallelism

In this case, the algorithm is broken into several different parts which run on different processors. Data flows through this network of processors and is acted upon rather like a production line. All of the data need not necessarily pass through every processor but this is frequently the case. This is often the most difficult sort of parallelism to program efficiently. The usual configuration is such that one master processor, which deals with input and output, controls the passage of data to several slave processors. These slave processors may be linked in a way most appropriate to the algorithm, but must also facilitate the passage of data between slave and master. The passage of data will also be interspersed with the passage of messages from the master to control the behaviour of the slaves. This usually necessitates a reasonably complex message passing system. The major obstacle to efficiency is load-balancing; if one processor dominates the execution time it becomes a bottleneck.

5.2 Practical Applications

The nature of the problem to be solved dictates the parallelism that is used. A problem may be best solved with a combination of different kinds of parallelism.

5.2.1 MD8 - A Parallel Implementation of PROMDL from GROMOS87

The theory and use of protein dynamics and energy minimisation have been covered in chapter 4. The major limitation to the use of these methods is the

amount of computation time required. If there are n atoms in the system, the number of calculations to describe the covalent interactions is proportional to n. However, the number of calculations needed to describe completely the non-covalent interactions is proportional to n^2. For a large number of atoms the calculation of all non-bonded interactions becomes prohibitive even for supercomputers. Therefore, in order to reduce the number of calculations, distance cut-off techniques are used. It is assumed that charges in proteins are never monopoles, thus the Coulombic interaction is inversely proportional to the cube of the atomic separation (eqn 4.4). The Coulombic interaction energy between atoms therefore drops off rapidly as they are separated. The non-bonded interaction need only be calculated over a short distance around each atom, typically 10 Å. In practice, for a given atom, interactions are only considered with other atoms that are within a specified radius of that atom. A list of these atoms is stored for every atom. This list is updated every few steps, often 10. Between updates the list is used to determine which calculations are performed. Thus the number of non-bonded calculations is proportional to kn, where k is the average number of atoms in a sphere of volume $\frac{4}{3}\pi r^3$. Only when the list is updated are the distances between all atoms calculated. In this way the number of calculations is kept to a manageble level for proteins *in vacuo*. When a large number of solvent molecules is included the number of calculations becomes very large, as k is in the order of 100 for a sphere of radius 10 Å. Therefore, energy minimisation and dynamics are usually carried out *in vacuo*, which often does not produce satisfactory results. The programs PROEML and PROMDL, from the program suite GROMOS87, were used in molecular modelling (see chapter 4) to optimise model structures. Minimisations and simulations of single proteins (approximately 2000 atoms) *in vacuo* were feasible using a VAX 11/750. However, when large numbers of solvent molecules were added (>3000) the CPU time taken was too long to be of practical use.

Molecular Simulation Programs

There are many programs written to perform macromolecular energy calculations: CHARMM, AMBER (Weiner and Kollman, 1981), GROMOS87 and X-PLOR, to name but a few. All carry out similar functions; energy minimisation using different algorithms, molecular dynamics, coordinate manipulation, and analysis of results. The input to these programs is also very similar:

- Atomic coordinates
- Molecular topology of these coordinates
- Bonded and non-bonded parameters
- Control parameters

The atomic coordinates descibe the cartesian position xyz for each atom in the system. The format of this information usually varies from program to program. The connectivity of these atoms must be defined in some way. This is usually done by listing atom pairs involved in covalent bonds, atom triplets involved in covalent angles, and atom quartets involved in dihedral angles. A list is also constructed of those atoms that cannot interact non-covalently. Obviously this description of molecular topology constitutes a large amount of information. The parameters that define the force-field must also be defined. This information is sometimes combined with the molecular topology or alternatively as separate files. The energy constants and optimum bond lengths and angles must be defined for each combination of atoms. The Lennnard-Jones and Coulombic parameters must also be defined for all combinations of atom pairs. Again this represents a large library of information, which is usually restricted to those atoms found commonly in proteins or nucleic acids. This information describes the potential energy of the system. Control parameters are needed to describe the way the minimisation or dynamics are carried out.

Algorithms

The algorithms for both energy minimisation and molecular dynamics share the same basic organisation (figure 5–4). Both require the force on each atom to be calculated, but they differ in what is done with this information. Therefore, these algorithms are based around subroutines that calculate this force and the potential energy associated with it. Within these subroutines most computing time is spent in the calculation of the non-bonded force terms, for reasons outlined above. Therefore, the obvious target for parallelisation is this non-bonded force calculation, especially if the aim is to study highly solvated proteins. Since parallel computing has become possible different algorithms for molecular dynamics calculations have been implemented. Often these have been applied to custom built architectures but have since been applied to off-the-shelf parallel machines. The sort of algorithm that can be utilised depends on the system to be studied. Molecular dynamics simulations of homogeneous systems, often thousands of point atoms, require only local information for each particle. This problem therefore lends itself to geometric parallelism. Simulation of heterogenous systems such as proteins is more suited to algorithmic parallelism, due to the long range forces that must be calculated. A parallel algorithm, the systolic loop, was reported in the mid-1980's as a method for calculating long range interactions using a MIMD architecture (Ostlund and Whiteside, 1984). This algorithm has been successfully applied to transputer based MIMD systems for protein dynamics and minimisation (Raine *et al.*, 1989; Heller *et al.*, 1990). A pipeline architecture has also been used to simulate gravitational many-body problems with long range forces (Sugimoto *et al.*, 1990).

Programming objectives

The use of GROMOS87 in the previously described molecular modelling made this program the basis for a parallel program. Therefore, the parallel version had to be compatible with the GROMOS87 file formats. The parallel machine used was the Edinburgh Meiko Computing Surface. Some specific aspects of the

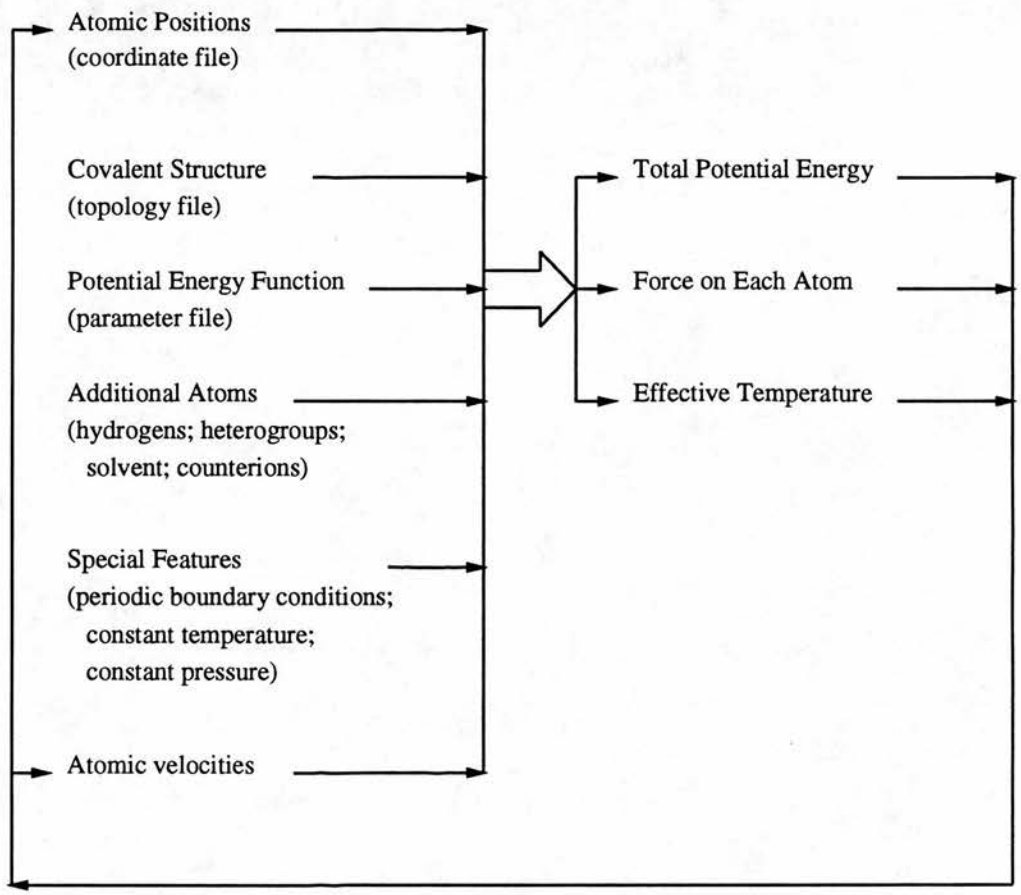


Figure 5–4: General method for both energy minimisation and molecular dynamics calculations.

implementation were felt to be very important, thus dictating the final form of the code:

- Use of the GROMOS87 forcefield.
- Use of standard GROMOS87 file formats.
- Implementation of the SHAKE algorithm.
- Fortran-77 code.
- Simple, asynchronous, anarchic communication.

The first three points were chosen for reasons of compatibility with previous work carried out with GROMOS87. The use of Fortran-77 was desirable to aid portability to any other MIMD machine. One of the main objectives was to make the parallel implementation as simple as possible, making the code easier to write and aiding portability. A simple parallelism had also to include simple communications between processors. In the time given it was not possible to achieve all of these initial aims. At the start of program development the CS did not fully support high level message passing from either Fortran or C. The only comprehensive parallel programming language implemented was Occam. It was possible to use Fortran subroutines from within Occam code using simple Fortran subroutine calls for communication to and from the Occam. It was therefore possible to write a Fortran master process to deal with input and output and some of the calculations. However, the slave processes which carry out the non-bonded and bonded force calculations were written in Occam. The slaves and masters communicate using an Occam harness. Since writing the code parallel, communication has become possible from within Fortran using CSTools, making Occam redundant.

Architecture

If we assume that all pairwise forces are to be calculated then every atom i must see all other atoms j provided $i \neq j$. This number of calculations can be halved

by using Newton's third law, of equal and opposite forces. Therefore, forces are calculated between every atom i and all other atoms j such that $i < j$. Every atom must see all other atoms at least once. The n atoms i can be distributed over x processors with integer $\frac{n}{x}$ atoms per processor. The n atoms j can be passed one by one to each processor and the force calculated between the atoms provided $i < j$. The simplest arrangement of processors for this operation is a pipeline into which atoms are pushed, being passed from one processor to the next. Two different communications strategies were considered. In the first case atoms were passed around a continuous pipeline (ring). Forces were accumulated by both stationary atoms and moving atoms. The moving atoms finally pass out of the end of the pipeline and are collected; the stationary atoms then leave the pipeline and are collected (figure 5–5). In the second scheme an atom passed to a processor is replicated and passed on. The stationary atoms collect forces; the moving atoms with forces pass out of the pipeline by a second communications line to be collected. Once all moving atoms have passed through a processor the static atoms leave and are collected (figure 5–6). The obvious disadvantage with the second strategy is that the amount of information leaving the pipeline is x times greater, where x is the number of processors. This creates a large communications bottleneck. The alternative was to design some system to decrease the amount of information produced, but this would have required a more complex communications strategy, in opposition to one of the main initial aims. Therefore, the first, continuous pipeline (ring) architecture was used as no data replication was needed using this scheme.

Comparison to other systolic methods

Although the method employed in the parallel implementation of PROMDL presented here has features of a systolic loop it also has many of the characteristics of a pipeline. The major difference when compared to other methods is the distribution of data prior to the force calculation. In the other systolic loop methods outlined below the data items between which forces are calculated, generally a static set of atoms and a moving set of atoms, are all

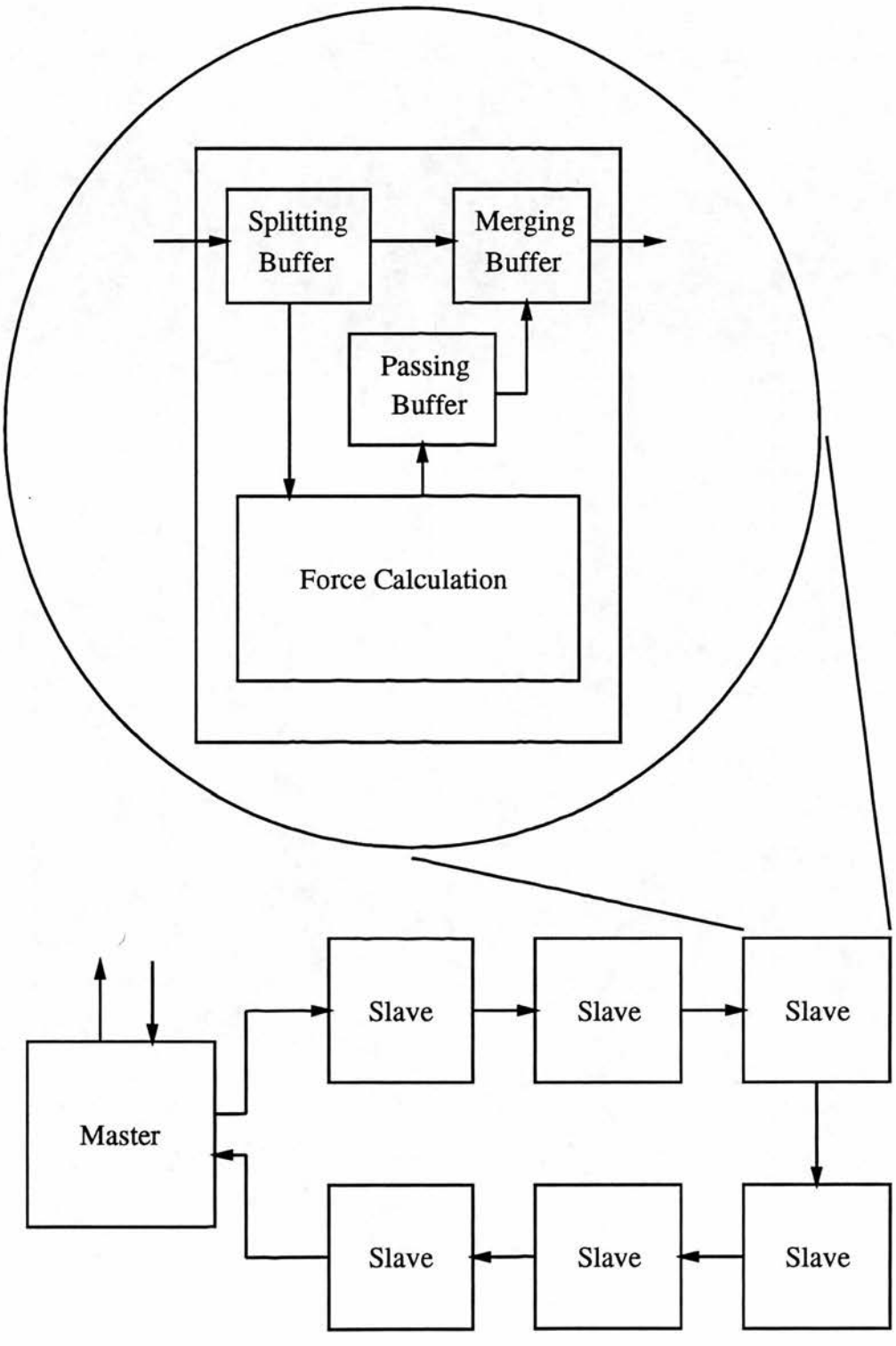


Figure 5–5: Ring topology for parallel non-bonded force calculation.

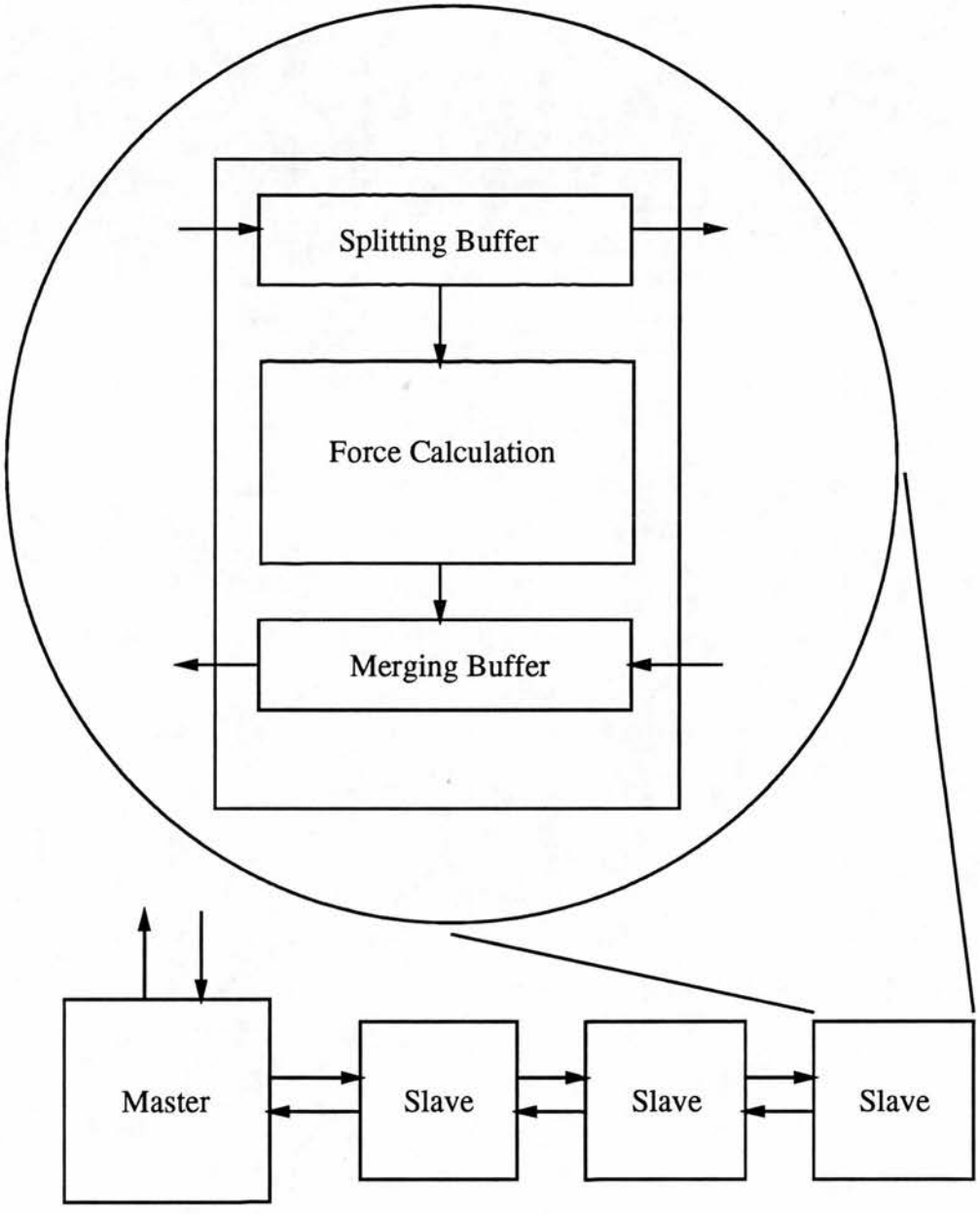


Figure 5–6: Pipe topology for parallel non-bonded force calculation.

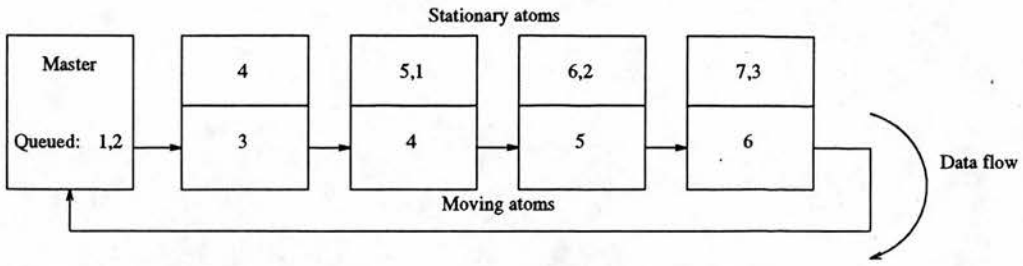


Figure 5–7: Data flow in the MD8 algorithm, for 7 particles.

distributed across the P processors before calculation begins. In the MD8 method the static atoms are distibuted prior to calculation but the moving atoms are passed through the static atoms in a pipelined manner (figure 5–7). Therefore, all moving atoms are not present in the ring at one time unless the number of atoms equals the number of processors.

A classical systolic method has been described by at least two groups (Raine *et al.*, 1989; Brugè *et al.*, 1988). Groups of atoms are distributed around a ring of P processors (figure 5–8). These represent the static atoms which do not move during calculation. Copies of each group are made - the moving atoms. The force between the static set and its copy are first calculated. Then the moving atoms on each processor are passed to the neighbouring processor (a pulse of movement involving all processors). After P such pulses the moving atoms are returned to their home processor. Using Newton's third law (as outlined above) the forces between all atoms are calculated after only $(P-1)/2$ pulses, provided the moving atoms have some way of accumulating forces. The remaining $(P+1)/2$ pulses are pure communication to return moving atoms to their home processors. This method is used in the systolic loop double method - SLD (Raine *et al.*, 1989). However, it is possible to divide the calculation of forces such that calculations are performed at each of the $(P-1)$ pulses, leaving only one pulse of pure communication to return moving atoms to their homes (Brugè *et al.*, 1988). It has been suggested that such a method is generally applicable for odd and even values of P and non-evenly distributed particles (Brugè and Fornili, 1991).

The requirement that moving atoms return their accumulated forces to their

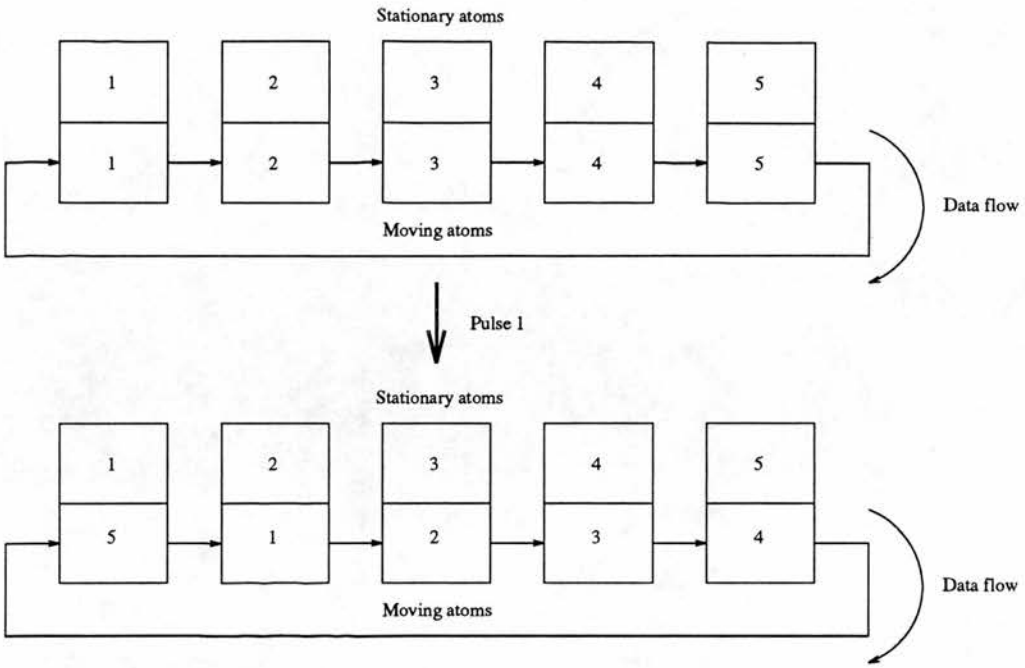


Figure 5–8: The Systolic Loop Double method for 5 particles.

home processors imposes a minimum of P communication pulses for a cycle to be complete. However, if forces can be returned to home processors during the $(P-1)/2$ pulses then particles need not return to their home processors. Such a method has been implemented using a twin communications ring (the program EGO, Heller *et al.*, 1990). As with the SLD method outlined above, particles are distibuted around the ring then copied to give moving sets. At each pulse forces are calculated then passed into a second ring which carries them back to their home processors in the opposite direction to the movement of particles (figure 5–9). In this way, after $(P-1)/2$ pulses all interactions have been calculated and the moving atoms can be 'deleted'. The amount of communication compared to the SLD method is approximately half. Also, the communication of force packets can be overlapped with the communication of coordinate packets and the calculation of new forces. The EGO method also incorporates a node specifically for the calculation of explicit hydrogen bonds between atoms.

A different systolic method has been suggested (Raine *et al.*, 1989), which uses only one data packet per particle - systolic loop single (SLS). The interaction between particles is evaluated in 'evaluator' processes, each of which is home to

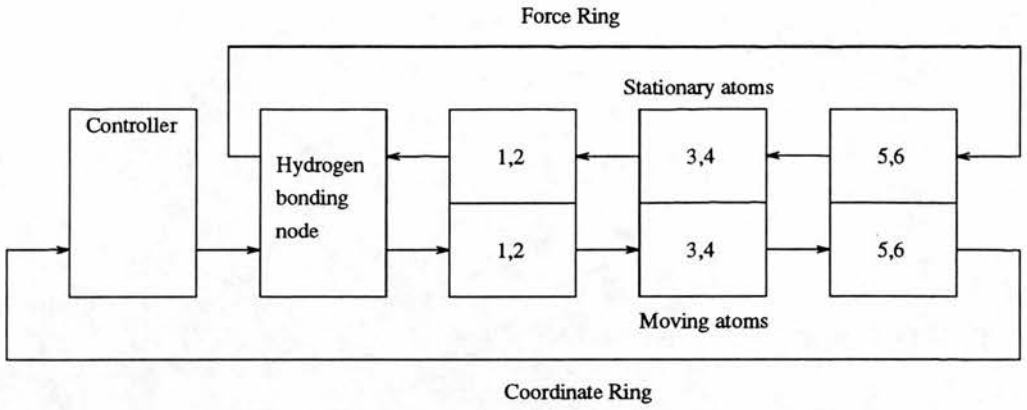


Figure 5–9: The topology of communication used in the program EGO, for 6 particles.

two particles. A 'head' process is home to one particle, and a 'tail' process is required to turn data items around as they pass through the 'ring' (figure 5–10). After N pulses all pairwise interactions have been calculated and all particles are returned to their home processors. This method does not require a copying of particle data to start the systolic loop. It also does not require force arrays to be held on home processors, instead they circulate with each particle. The method has the advantage over the SLD method that for a given (odd) number of particles, N, only $(N-1)/2$ evaluator processes, as opposed to N processes are needed. The method can also be extended to overlap the communication of force accumulators with the calculation of forces. This necessitates a one step lag between a particle data packet and its force packet (Raine *et al.*, 1989).

All of the methods outlined above can be adapted to use groups of particles as data items rather than individual particles. This method minimises the time associated with communications startup (which is a constant value irrespective of the size of the data packet). In addition, the PROMDL program can construct a nonbonded pairlist on the basis of the distance between charge groups rather than individual atoms. A charge group is defined as a collection of atoms whose partial charges sum to zero or an integer multiple of the charge on one electron. The division of particles into collections of charge groups therefore would

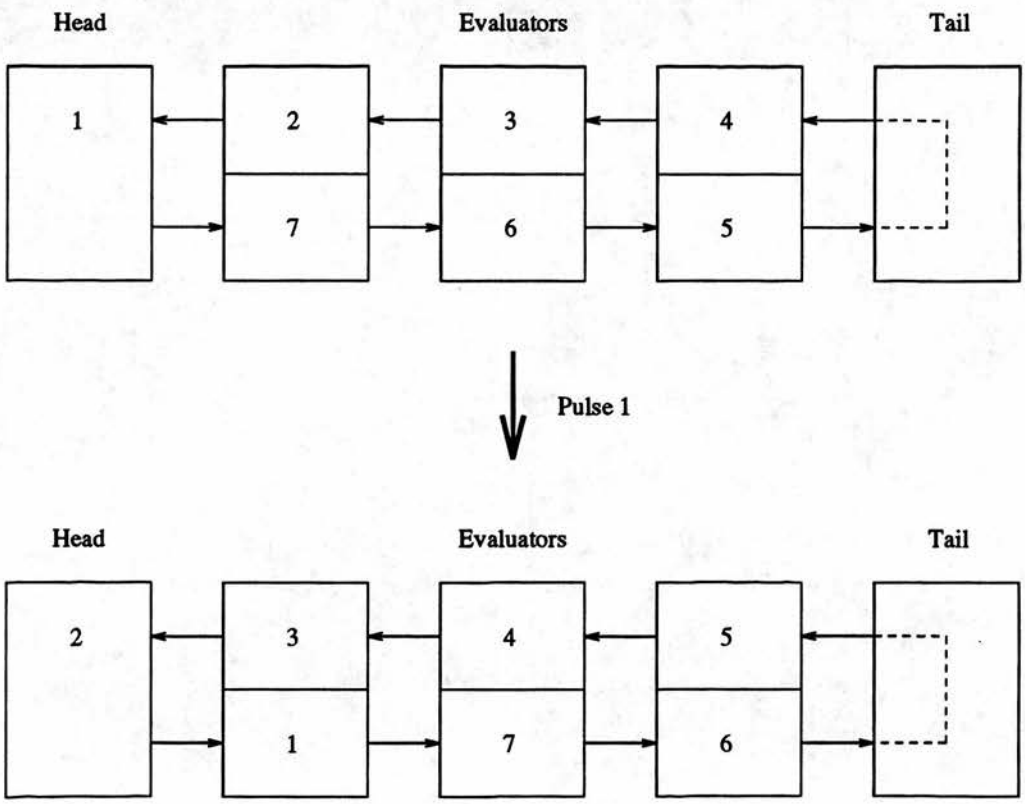


Figure 5–10: The Systolic Loop Single method for 7 particles.

minimise the communication startup overhead and allow charge group based pairlist construction.

The above methods must also incorporate the calculation of covalent as well as non-bonded force terms for proteins. In general the atoms involved in covalent terms are local; they are within a particle group or span two particle groups. The information needed to calculate the covalent force terms can be easily gathered by some small amount of communication between neighbouring processors. This method is employed in both EGO and the protein version of the SLS algorithm (Raine, 1991). A similar method could have been used in the MD8 program but was not due to ease of programming a separate ring for the bonded force calculation.

Memory

Memory is only local to each processor on the CS. In general each processor has 4 MBytes of RAM memory, with 4 Kbytes of on chip memory. The major storage requirement of the non-bonded force calculation is the list of atoms between which forces are to be calculated - the non-bonded pair list. For each atom j, information about all atoms $i < j$ which are within the cutoff range must be stored. If there are n atoms and x processors there will be $\frac{n}{x}$ atoms per processor. There will be approximately 200 atoms within a cutoff range of 10 Å with $i < j$. Therefore the non-bonded pair list must be an array of at least $\frac{n}{x}$ by 200. To allow for larger cutoff ranges, space for 500 atoms per atom j was allocated. To maintain compatibility with GROMOS87 a maximum of 15000 atoms was implemented. Assuming a minimum of 12 processors for such a large system, the non-bonded pair list array was 1250 by 500. Using Occam INT32 variables to store this list required 2.5MBytes. The use of Occam INT16 variables would have halved the size of this array but was not implemented. The remaining memory was mostly consumed by arrays describing topology and non-bonded interaction parameters.

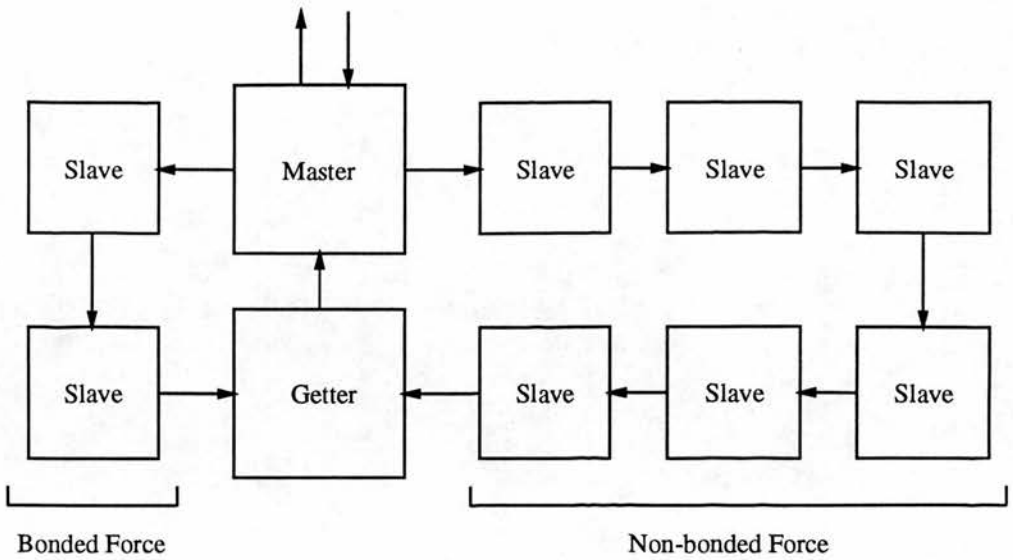


Figure 5–11: Final topology for complete parallel force calculation.

Message passing

The passage of data from processor to processor was designed to be very simple. Each data packet contained the three coordinates for an atom, an identification number, atom type, atom charge, group number, and the three collected forces. The first word in an atom packet was a flag determining the action of the slave processor. The second word was the number of the destination processor for stationary atoms. Each atom packet was therefore 12, 32 bit words long thus occupying 48 bytes.

The Bonded Force Calculation

For protein simulations *in vacuo*, the calculation of the bonded forces is not insignificant, unless all non-bonded interactions are calculated: this fact will be demonstrated later. For highly solvated systems no bonded forces are calculated for solvent molecules, therefore the non-bonded calculation predominates. Therefore, the bonded force calculation is carried out in parallel with the non-bonded force calculation. This is carried out by a second ring of processors, independent of the non-bonded ring (figure 5–11).

Inclusion of Solvent Molecules

Solvent was included explicitly as complete molecules, rather than implicitly as a component of the force field. Computational time was saved by not calculating the bonded interactions for these solvent molecules. Rather their covalent geometry was maintained using the SHAKE algorithm. They only participate in non-bonded interactions - the reason that they are simulated at all. After each integration of the equations of motion the solvent molecules were constrained using SHAKE, consuming much less time than calculating all bond and angle forces explicitly.

The Twin Range Method

This cutoff method is used to reduce the number of non-bonded force calculations carried out. As the name suggests two cutoff distances are used: for example a short range (10 Å) and a long range distance (15 Å). The short range distance is used to calculate the non-bonded pair-list and hence the short range forces at every step. The long range forces are calculated when the pair-list is set up. These are the forces between an atom and other atoms in the shell between the short range and long range distance (figure 5–12). These forces are collected from the non-bonded ring and stored on the process responsible for collection of all forces: the "getter" process. At subsequent steps using the pair-list, which calculate the short range forces, these long range forces are added to the short range forces by the getter process before passing the force array back to the master process. This technique makes the assumption that the atoms in the long range shell around an atom move only a small distance between non-bonded pair-list (long range force) calculation. This assumption is valid provided the number of steps between pair-list update is not large and that the short and long range cutoffs are of a large enough value.

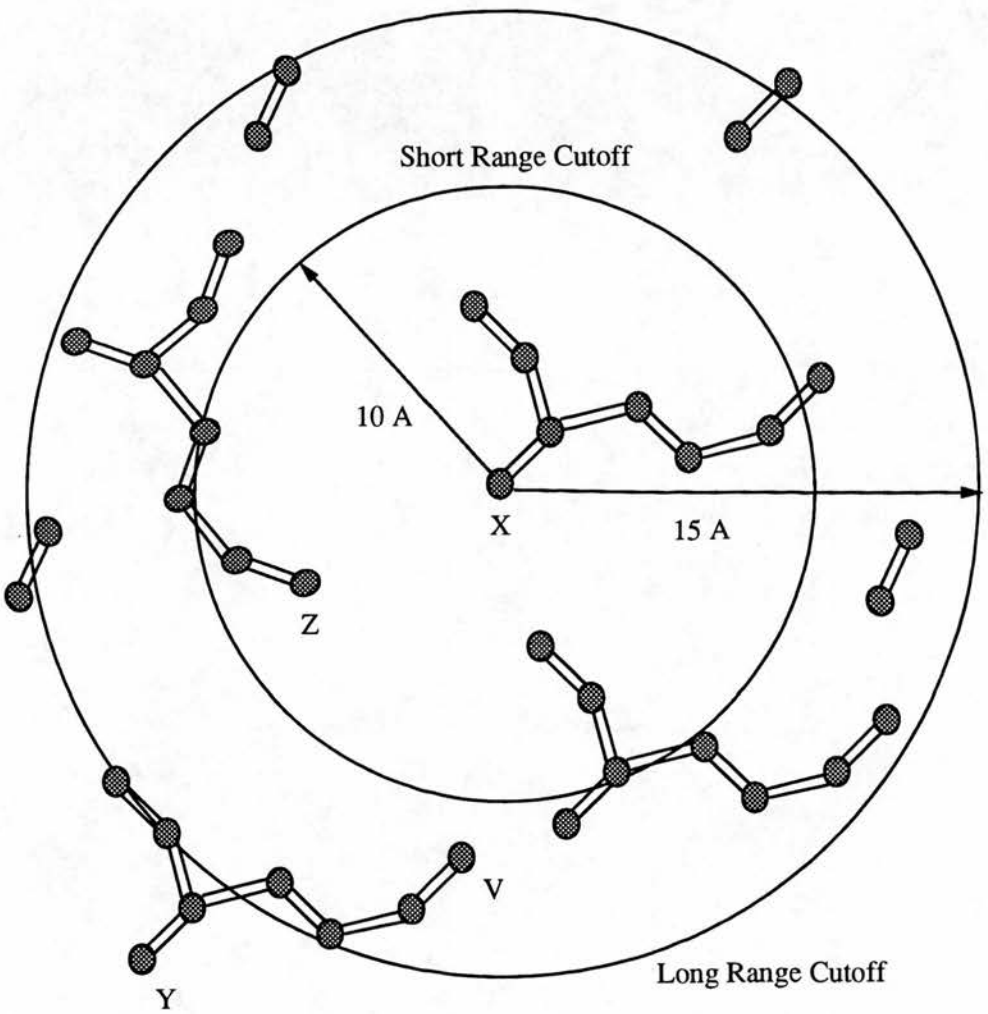


Figure 5–12: The twin range method in non-bonded force calculations. The interaction between atom X and Z is calculated every timestep, between X and V only when the pair-list is updated, and between X and Y not at all.

Temperature Control

The temperature of systems under simulation may be controlled using the coupled water bath technique (Berendsen *et al.*, 1984). Velocities are scaled by a factor λ,

$$\lambda = \left[1 + \frac{\delta t}{\tau_T}\left(\frac{T_0}{T} - 1\right)\right]^{\frac{1}{2}} \tag{5.1}$$

where τ is the coupling constant, T_0 is the reference temperature, and T is the present temperature of the system. Tests with homogeneous systems have suggested that a coupling constant of 0.1 or higher is preferable for reliable dynamic properties (Berendsen *et al.*, 1984).

Configuration

The connectivity between processors is explicitly defined within the Occam program. This placement was constructed in such a way that the number of bonded or non-bonded processors could be changed easily by altering 2 program parameters, BPROC and NBPROC, respectively. The executable code could then be reconfigured at run time. A library of files that described the physical links between transputers was created using an automatic wiring tool. The transputers were wired with the appropriate wiring file before the program was executed. It would have been possible to configure the processors dynamically at run time, by running the wiring utility prior to execution. This option was not taken because of the time needed to create the wiring information each time, which increases dramatically as the number of processors increases.

Timing Results

The parallel program (MD8) was tested with three different sized systems; crambin (396 atoms), MUP (1597 atoms), and MUP plus many water molecules (10366 atoms). The time taken per integration cycle was calculated from the

		Time per step (seconds)		
	Atoms	VAX 11/750	9 x T800	54 x T800
Crambin	396	16	0.7 (23)	0.57 (28)
MUP	1597	75	4 (19)	2.0 (37.5)
MUP + Water	10366	900	94 (9.5)	30 (30)

Table 5–1: Time taken for one integration step with different numbers of atoms on a VAX 11/750, 9 and 54 T800 transputers. Figures in brackets indicate the speed-up factor compared to the VAX 11/750.

time taken to complete all steps divided by the number of integration steps. This was compared with the time taken to carry out the same calculations using the serial version of GROMOS87 on a VAX 11/750. The timings were repeated with different numbers of processors in the non-bonded ring using first one then two bonded processors. The results are summarized in table 5–1 and presented in detail in figures 5–13, 5–14, and 5–15. It can be seen that it was possible to achieve a maximal speed-up factor of between 30 and 40 fold with respect to the VAX 11/750 timings. It can also be seen that the bonded force becomes significant in the case of crambin and MUP *in vacuo* as the time taken for the non-bonded calculation decreases. This is not observed for MUP plus solvent because the non-bonded calculation involves both protein and water atoms whilst the bonded calculation only involves the protein atoms.

Efficiency

If the timings are inverted, converting them to cycles per second and plotted against the number of non-bonded processors the scalability of the program can be assessed. For the ideal parallel program perfect scalability would be observed: the number of cycles per second doubles as the number of processors doubles. The scalability of this algorithm with MUP plus water is far from ideal (figure 5–16). The number of cycles per second falls off as the number of processors increases. An algorithm which behaves in this way has been termed mildly scalable (Fincham and Mitchell, 1991). This indicates that although the algorithm does scale, because the number of cycles per second increases, the increase is not proportional to the increase in the number of processors. In

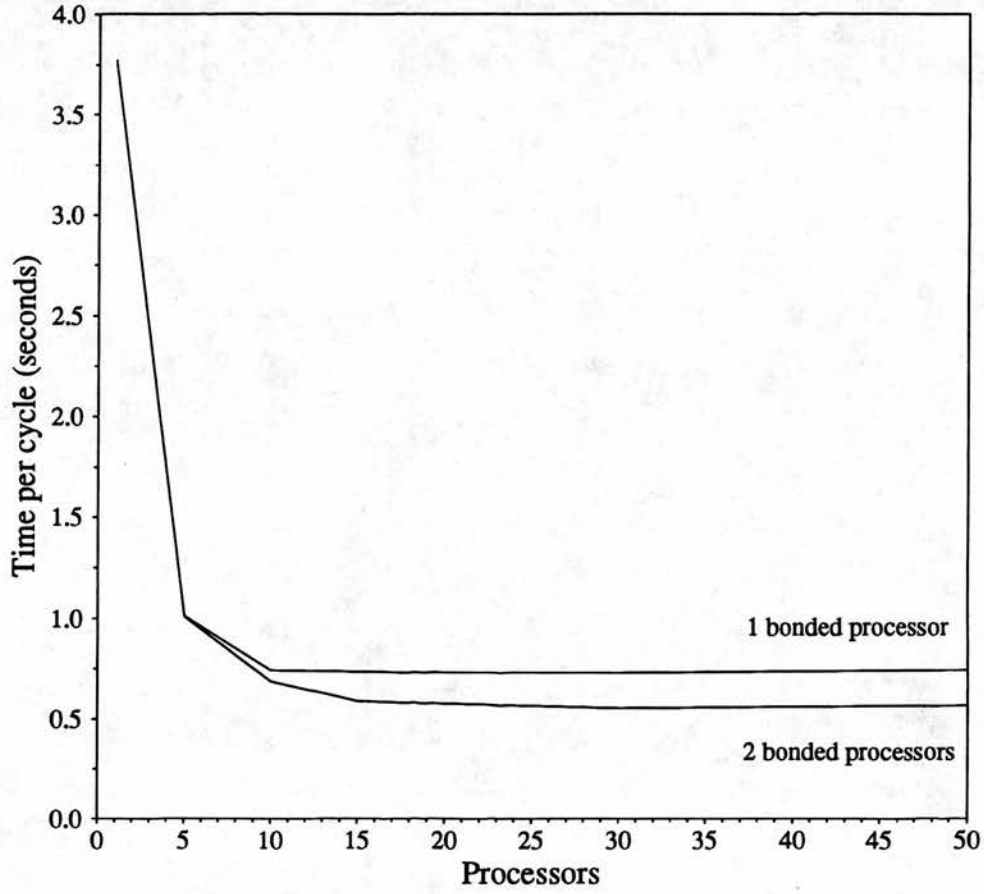


Figure 5–13: Time per cycle for MD8 with different numbers of non-bonded and bonded processors for crambin.

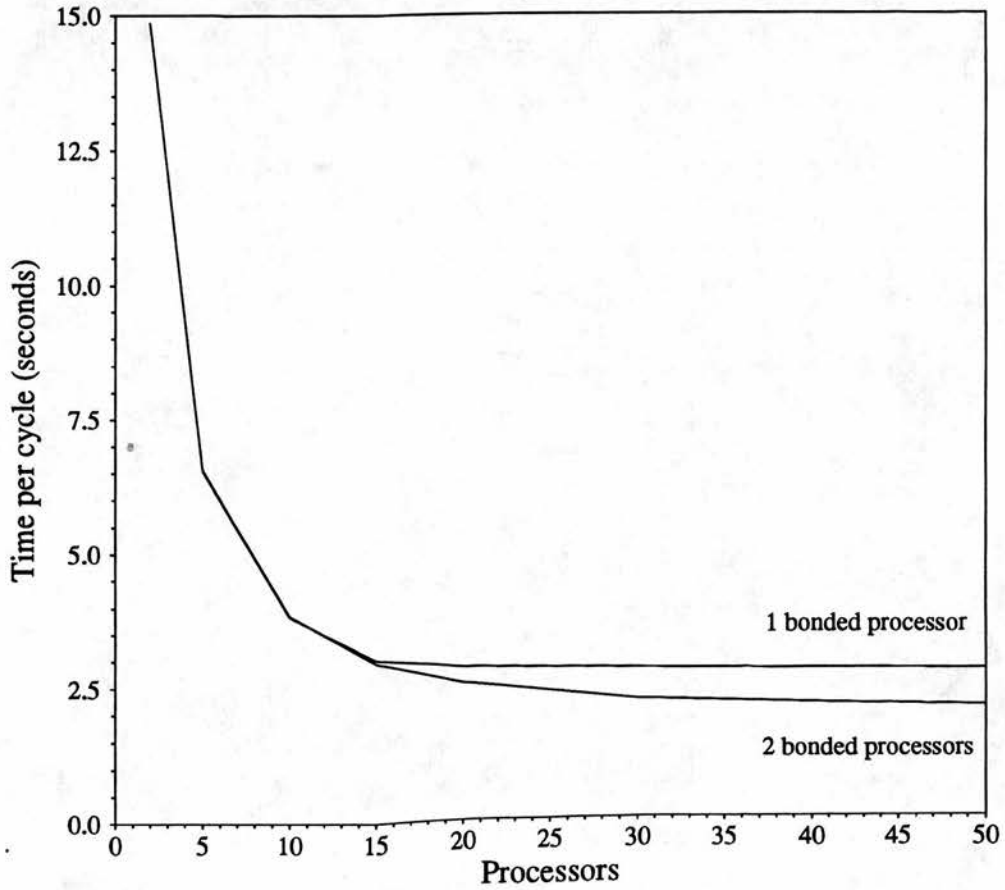


Figure 5–14: Time per cycle for MD8 with different numbers of non-bonded and bonded processors for MUP.

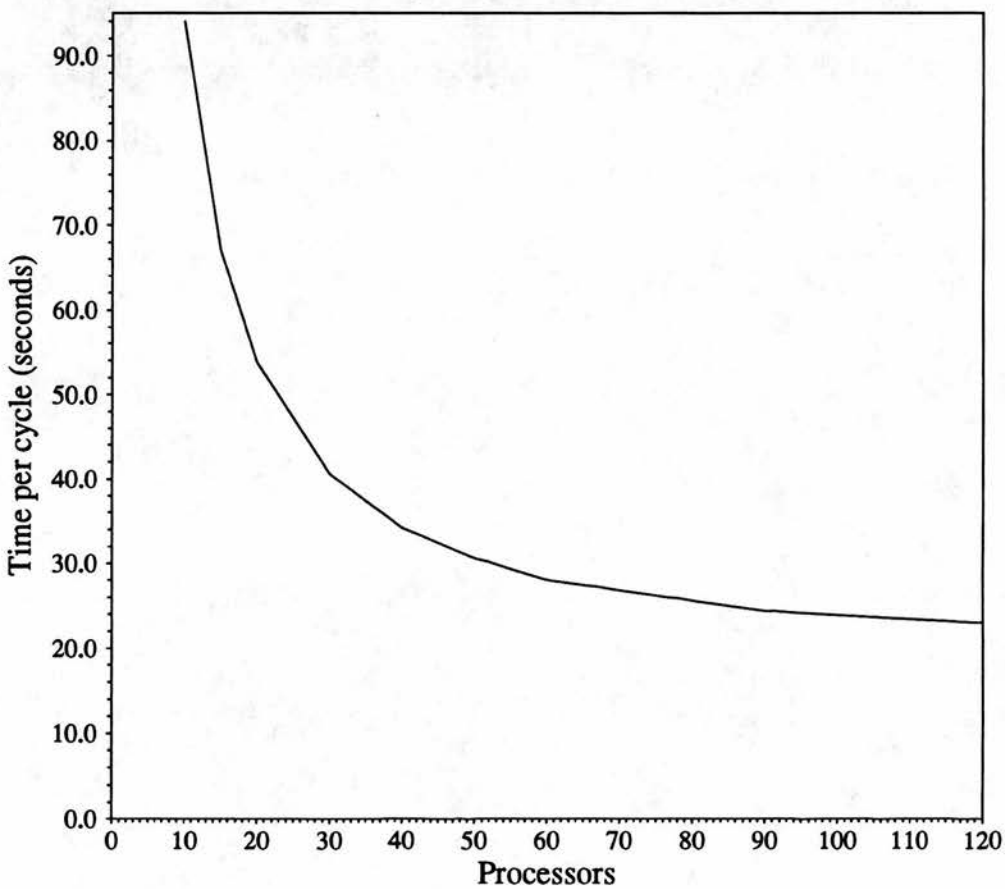


Figure 5–15: Time per cycle for MD8 with different numbers of non-bonded processors for MUP plus solvent.

addition the increase in the number of cycles decreases as the absolute number of processors increases. This behaviour suggests some bottleneck as the number of processors increases, in this case the rate determining step is the speed of communication between processors.

5.2.2 Crystallographic Refinement

Three dimensional models of proteins based on X-ray diffraction data are derived initially from electron density maps. The atomic coordinates based on such models are not very accurate, due to limitations in resolution, initial phases, and inadequate interpretation of the electron density map. To understand the chemistry of these proteins and how they function, more accurate models are required. It is important, therefore, to improve the models of proteins by refining them to the fullest possible extent. Refinement refers to the process of adjusting the parameters of the models, usually the positions (x_j, y_j, z_j) and the thermal parameters (B_j) of the j atoms within the unit cell of the structure, in order to improve the agreement between the amplitudes of the observed reflections ($|F_o|$) and the values calculated from the model parameters ($|F_c|$). The agreement between observed and model amplitudes is usually assessed by the R-factor:

$$R = \frac{\sum ||F_o| - |F_c||}{\sum |F_o|} \tag{5.2}$$

The sums are over all observed reflections in a data set. Closer agreement between observed amplitudes $|F_o|$ and calculated values $|F_c|$ is reflected by a lower value of R, and a better model. Some refinement techniques do not use the agreement between reciprocal space data, rather a real space agreement is used (Diamond, 1985). This technique monitors the agreement between the observed electron density (ρ_o) and the model electron density (ρ_m).

Different methods for refining model parameters to minimise the R-factor have been implemented. The general method employed for electron density maps derived from heavy atom isomorphous replacement requires much manual

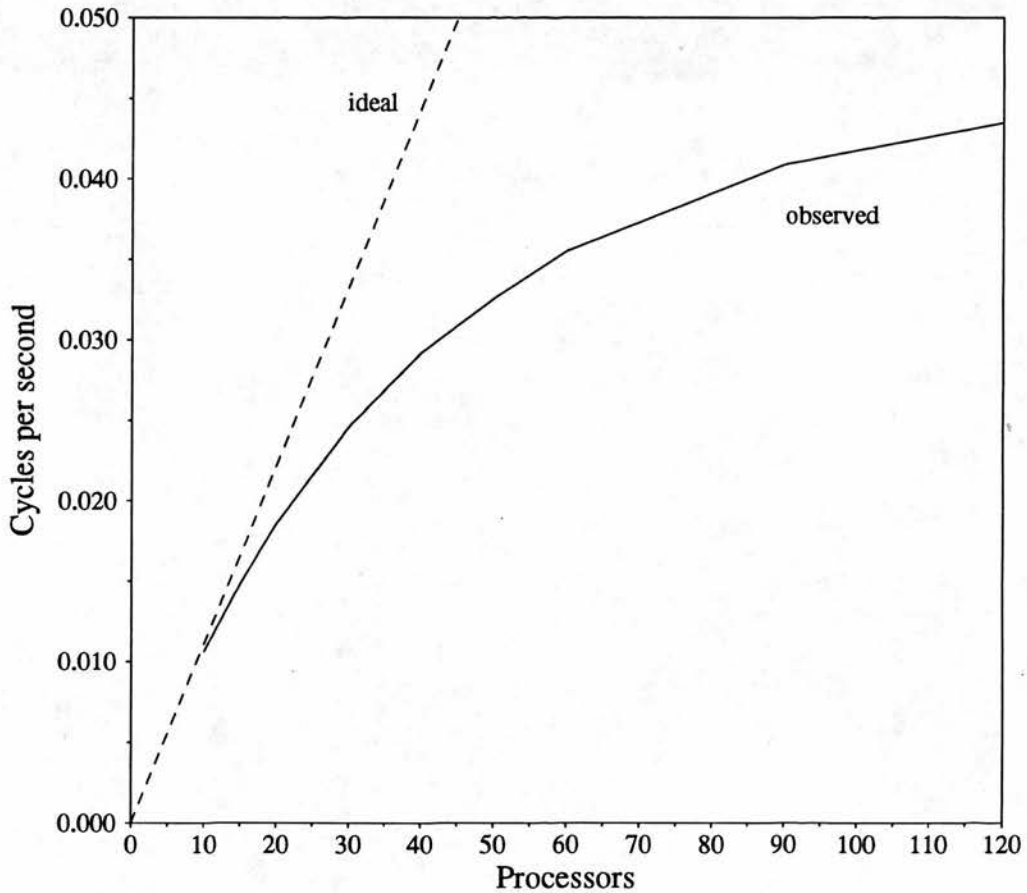


Figure 5–16: Scalability of the MD8 program with respect to number of processors for MUP plus solvent.

intervention. An initial electron density map is used to build a model. Phases are derived from the Fourier transform of this model. These phases are then combined with the observed amplitudes to produce a new electron density map. The model can then be moved to optimise agreement with this map and the whole cycle is repeated. This conventional refinement usually consists of several cycles of computational, automated refinement of model parameters followed by manual rebuilding of the model using interactive graphics. The whole process is then repeated in a cyclic manner until no further improvement in R-factor is achieved.

Stereochemically Restrained Least-Squares Crystallographic Refinement

In macromolecular crystallography there is usually a paucity of observed data, due to the relatively weak diffraction obtained from macromolecular crystals. Hence the number of observations (reflections) may be of the same order as the number of model parameters (atoms) and the problem is therefore under-determined. This can be compared to the small molecule case where the number of observations can often be 10 times greater than the number of atoms, hence the problem is over-determined. In the macromolecular case the diffraction data may be supplemented with prior knowledge about stereochemistry. This technique effectively increases the number of observations (Hendrickson, 1985). These stereochemical conditions serve to restrict model features to a realistic range of possibilities. These conditions are termed restraints in contrast to constraints, which confine model features to specific values.

The program PROLSQ includes stereochemical information in a least squares refinement of protein structure (Konnert and Hendrickson, 1980). The process of least squares optimization of a function has been covered previously (chapter 4). Briefly, therefore, each piece of information is treated as an observational equation,

$$g_{obs} = g_{calc}(\mathbf{x}) + \epsilon \tag{5.3}$$

in which a residual ϵ describes the discrepancy between an observation g_{obs} and the theoretical value g_{calc} computed from the parameters $\mathbf{x}$ of a model. The observations are both crystallographic and stereochemical. The least-squares optimization procedure defines the "best" set of parameters as that which minimises the sum, over all observations, of the squared residuals:

$$\phi(\mathbf{x}) = \sum_h w_h [g_h^{obs} - g_h^{calc}(\mathbf{x})]^2 \tag{5.4}$$

where each observation h can be weighted by the appropriate value w_h. There are several different classes of observations of the form ϕ_i which become terms in a grand function for minimisation:

$$\Phi = \sum_i \phi_i \tag{5.5}$$

Twelve observational functions have been included in the program PROLSQ, descriptions of which are given elsewhere (Hendrickson, 1985). Briefly, functions are:

- Crystallographic
  - The difference between observed and calculated structure factors.
- Stereochemical
  - Bonding distances
  - Atom coplanarity within certain groups
  - Stereoconfiguration at chiral centers
  - Non-bonded contacts
  - Torsion angles

- Stereochemistry of atomic motion
  - Isotropic temperature factors
  - Anisotropic temperature factors
- Non-crystallographic symmetry
  - Atomic positions
  - Thermal parameters
- Other factors
  - Resistance to excessive parameter shifts
  - Occupancy factors

The standard version of PROLSQ calculates the theoretical values for diffraction information from the model using a space-group-optimized direct summation method. Stereochemical information is calculated from the atomic coordinates and compared to library values. The least squares normal matrix is set up from these observational terms, but is not inverted to determine a direct solution, rather a conjugate gradients minimisation technique is used to determine parameter shifts. Only block diagonal terms are used to reduce computational time, making the assumption that observational parameters are not highly correlated. The refinement process in general, and more particularly the structure factor calculation and setting up of the derivative matrix are CPU intensive (Hendrickson, 1985). This is particularly true for large numbers of atoms or reflections.

The PROLSQ program was the target for implementation on a parallel computer. Parallelism had already been applied to the PROLSQ program, using a Floating Point Systems AP 120B array processor (Hendrickson, 1985). More recently the work of Raftery *et al.* (to be published) implemented the PROLSQ program on an ICL Distributed Array Processor (DAP). Like the CM-200 the ICL DAP was a SIMD computer with a large number of processor elements

Machine	Host	Host time	Parallel time	Total time
ICL 2976	-	5500	-	5500
ICL DAP	ICL 2976	73	253	326
AMT DAP	Sun 3/60	147	132	279
AMT DAP	Sun 4/25	80	132	212
Sun 4/20	-	272	-	272
CM-200	Sun 4/370	32	2.9	34

Table 5–2: Timings in seconds for parallel implementations of PROLSQ for one cycle of refinement of BLG (1196 atoms, 3564 reflections in the space group $B22_12$).

(4096). These PEs could all perform the same operation in parallel on multiple data items. Unlike the CM-200, PEs were connected in a two-dimensional grid or toroidal topology. The implementation aimed to parallelise the most computationally intensive parts of the program, namely structure factor calculation from model coordinates and the conjugate gradients solution of the normal matrix. Parallelisation of structure factor calculation was over the number of atoms. The contributions from up to 4096 atoms to a single reflection were calculated simultaneously. Significant improvements in execution time were obtained relative to an ICL 2976. (table 5–2).

The ICL DAP has been superseded by the AMT DAP in recent years. The ICL DAP version was converted to run on an AMT DAP 610 by the author. This involved some changes to subroutine calls and the interaction between the host computer and the DAP. The ICL DAP was a memory extension of the ICL serial host. The AMT DAP is hosted by a Sun or VAX front-end, data has to be explicitly passed from the host to the DAP in a similar manner to the CM-200 and its front-end. The performance of the Sun 3/60 front-end and AMT DAP for a small test case is significantly better than the ICL pair (table 5–2). The same program, PROLSQ, was also implemented on the CM-200. It was not possible to transfer DAP code directly to the CM-200. The DAP implementation had been written to optimize use of the DAP hardware. The CM-200 implementation was therefore written *ab initio*. Initial tests with a serial version of the code running on a Sun 4/20 suggested that the main target for parallelisation should be the structure factor calculation routine CALC (table 5–3). This routine is called from the structure factor derivative calculation subroutine XSFREF and the

R-factor test subroutine RTEST. The calculation of structure factors from models coordinates was parallelised over the number of atoms in the same way as in the DAP implementation. For a given reflection **h**,

$$F(\mathbf{h}) = \sum_{j}^{atoms} f_j \exp[2\pi i(\mathbf{h}\mathbf{x}_j)] \tag{5.6}$$

where f_j is the atomic scattering factor for atom j with fractional atomic coordinates $\mathbf{x}_j$. The serial version of PROLSQ used lookup tables to calculate the atomic scattering factor for each atom. This proved inefficient to implement in parallel, as is generally the case for data parallel computers because non-local communication between processors is required. Therefore, the calculation of atomic scattering factors was carried out directly using a four term analytical expression (International Tables Volume IV, 1972):

$$f(\sin\theta/\lambda) = \sum_{i=1}^{4} a_i \exp(-b_i \sin^2\theta/\lambda^2) + c \tag{5.7}$$

The contribution from all atoms to a reflection **h** was calculated simultaneously. There was no limitation to the number of atoms considered at one time due to the underlying CM-200 microcode implementation of virtual processors. All other calculations remained on the host front-end but were optimized as much as the Sun f77 compiler would allow (using the -O3 option). Data were explicitly passed from the front-end to the CM-200 using a CM Fortran library call. The derivatives needed to calculate the normal matrix were accumulated on the CM-200 and only passed back to the front-end when all structure factor calculations were completed. The timing results show an improvement relative to both the ICL DAP and AMT DAP implementations (table 5–2). No attempt was made to parallelise other parts of the PROLSQ code. This was because other parts, including the conjugate gradients minimisation, involved large amounts of non-local data reference which would require much communication of data between essentially random pairs of processors. This kind of communication is very inefficient on SIMD parallel computers. It is possible that some speed-up

Subroutine	Sun 4/20	CM-200
disref	2.85	0.68
plnref	0.82	0.40
chiref	0.30	0.13
vdwref	4.53	1.29
torref	1.70	0.72
xsfref	206.50	10.61
cgsolv	6.06	4.01
rtest	35.32	2.54
total	271.64	33.92

Table 5–3: Timings for individual subroutines in PROLSQ on both a Sun 4/20 and CM-200. All times are in seconds, the total time includes the time taken by the main program.

may have been obtained from parallelisation of the conjugate gradients minimiser, but this would only have been of any significance for large data sets.

5.2.3 The Molecular Replacement Translation Function

The use of the molecular replacement technique in protein crystallography has become more common. The method relies on first finding the correct orientation for the search molecule in the unit cell of the unknown structure, then determining the correct translational parameters for this rotated search molecule (Rossmann, 1972). The program BRUTE uses a linear correlation coefficient to determine the translational parameters (Fuginaga and Read, 1987). As the name suggests, a brute force approach is used with the correlation coefficient being calculated at every point in the search space. The time taken to calculate structure factors is reduced using molecular scattering factors (Lipson and Cochran, 1957). This makes the calculation of all Fs at a grid point proportional to the product of the number of reflections and symmetry operations. The initial calculation of the molecular scattering factors is proportional to the product of the number of atoms, reflections and symmetry operations. However, the time taken for large grids or fine grid searches is large. The translation function is often very flat, peaks only being well resolved by fine sampling of the search space. The grid sampling required to obtain the correct solution is usually no more than 1 Å. Timings were made for three proteins; *Drosophila* alcohol

Protein	Refs.	Atoms	Symmetry	Grid	Time G	Time F	Total
ADH	3666	1981	$P2_1$	100x100	799	4942	5750
BLG	1130	1496	$B22_12$	56x68x5	745	8617	9370
BLG	1130	1496	$B22_12$	56x68x41	745	70660	71405[a]
PT	7559	2951	$P2_12_12_1$	50x80x1	4857	6384	11255
PT	7559	2951	$P2_12_12_1$	50x80x200	4857	1276800	1281657[a]
PT	16384	2951	$P2_12_12_1$	50x80x200	10527	2767442	2777969[a]

Table 5–4: Timings (in seconds) for BRUTE with different test data on a Sun 4/20 SLC. [a] Figures extrapolated from other timing results.

dehydrogenase (ADH), lattice Y BLG (BLG), and pertussis toxin (PT) (crystallographic data kindly supplied by E.J.Gordon, A.S.McAlpine, and M.A.Turner respectively). Search models were 3α,20β-hydroxysteroid dehydrogenase (HSD; Ghosh *et al.*, 1991), MUP, and *E. Coli* heat labile enterotoxin (LT; Sixma *et al.*, 1991) respectively. Initial timings were made using the serial VAX version of BRUTE modified slightly to run on a Sun 4/20 SLC (compiled with f77 -O3). The time taken to calculate the molecular scattering factors (G), the translation search itself over all grid points (F), and the total CPU time was measured (table 5–4).

Method

The serial version of BRUTE was obtained from CCP4 (Daresbury Labs., UK.). The original version of BRUTE (Fuginaga and Read, 1987) was written for an Floating Point System FPS164 attached processor which was a pipelined machine capable of operations on long arrays. The core molecular scattering factor calculation and subsequent structure factor calculations were parallelised. The direct structure factor calculation of the type used in BRUTE was parallelised by applying calculations to all reflections simultaneously.The calculation of the molecular scattering factors G_j for each reflection $\mathbf{h}$,

$$G_j(\mathbf{h}) = \sum_{k}^{\text{atoms}} f_k \exp 2\pi i(\mathbf{h}.\mathbf{x}_{jk}) \tag{5.8}$$

was parallelised over the number of reflections, where f_k is the atomic scattering factor for each atom, and $\mathbf{x}_{jk}$ is the coordinate for each atom after application of symmetry operator j. The calculation of the structure factors F,

$$F(\mathbf{h}) = \sum_{j}^{\text{nsym}} G_j(\mathbf{h}) \exp 2\pi i(\mathbf{h}.R_j\Delta) \tag{5.9}$$

was also parallelised over the number of reflections, where R_j is the rotation matrix for symmetry operator j, and Δ is a translational shift in the atomic coordinates.

The program can be run on any size of CM and should transfer directly to the next-generation CM, the CM-5. All calculations are 32-bit although extension to 64-bit is possible using compiler flags. The program runs under the Unix operating system and is written in Fortran-77 and Thinking Machines CM-Fortran which is a subset of standard Fortran-90. The program is compiled and runs under the slicewise CM-Fortran execution model (section 5.1.4).

The program was modified to calculate a maximum of 1048576 grid points using up to 16,384 atoms and 16,384 reflections. These limits were not determined by the machine and therefore can be easily increased by changing parameter statements. Other changes to the original version of BRUTE were made. The program was modified to read X-PLOR formatted reflection files and standard Brookhaven coordinate files. The resolution limits for reflections were defined in Å and the writing of map sections to the output file was under user control. The parallelisation of the program was solely within the calculation of molecular scattering factors and the translation function calculation. The reading of reflections and atomic information remained serial - this data was explicitly passed to the CM. As with PROLSQ, the direct calculation of atomic scattering factors was implemented.

Results

The program was tested on a 16K processor CM-200 with 512 Weitek floating point vector units, and a total of 0.5 Gbytes of memory. Timings were made

Protein	Refs.	Atoms	Symmetry	Grid	Time G	Time F	Total
ADH	3673	1981	$P2_1$	100x100	1.84	15.95	34.1^a
ADH	3673	1981	$P2_1$	100x100	0.99	11.38	34.3^b
BLG	1157	1225	$B22_12$	56x68x41	2.47	388.1	658.2^a
BLG	1157	1225	$B22_12$	56x68x41	1.44	254.3	334.4^b
PT	7555	2951	$P2_12_12_1$	50x80x200	5.47	2018.5	2434.2^b
PT	16384	2951	$P2_12_12_1$	50x80x200	10.49	3477.6	3930.7^b

Table 5–5: Timings (in seconds) for BRUTE-CM with different test data using [a]one and [b]two CM-200 sequencers.

Protein	Reflections	8K-CM	16K-CM
ADH	3673	168	168
BLG	1157	108	214
PT	7555	-	527
PT	16384	-	707

Table 5–6: Speed-up of BRUTE-CM, running on one and two sequencers, compared to serial code on Sun 4/20 SLC.

using the full (16K) and half (8K) machine. Timing results and job parameters are shown in table 5–5. The speed-up factors for each test case are shown in table 5–6. The increase in performance was problem dependant; the larger the number of reflections, the better the speed-up. This was expected as the virtual processor ratio optimum for a CM hosted by a Sun is 32 and above. This ratio was reached on the whole machine only when 16384 reflections were used. It is probable that the speed-up factor of 700 obtained for PT with 16384 reflections represents the maximum speed-up possible. The minimum speed-up factor, observed for ADH, is artificially low because the time taken for the serial part of the program begins to become dominant.

5.2.4 Discussion

All three applications presented above were successful to varying degrees. The best improvements were obtained with SIMD parallelism. This does not mean that MIMD parallelism is inherently inefficient. Consideration must be made of the particular algorithm under investigation. The direct summation methods used in both PROLSQ and BRUTE lend themselves very well to SIMD parallelism. The same algorithms can also be parallelised using MIMD methods -

with several reflections being calculated on one processor. However, N-body algorithms such as those used in molecular dynamics simulations are inherently difficult to parallelise. The long-range interaction between particles neccessitates the movement of data from one proccessing element to another - often across many communication links. This need for communication is a problem for both MIMD and SIMD parallelism. A further problem for SIMD parallelism is the need for each processor to carry out the same number of operations on each data element. Therefore, the time saving algorithms generally used in molecular dynamics simulations, such as non-bonded pair lists, cannot be readily used.

The implementations presented above were optimised to varying degrees. The implementation of BRUTE on the CM was optimised as far as possible, without resorting to the use of lower level machine code. It is unlikely that the implementation of PROLSQ on the CM could have been optimised much further. The molecular dynamics algorithm implemented on the Meiko Computing Surface was not optimised. The transputer is a RISC architecture processor and therefore compiler optimisation of the code is possible. It is not clear whether the compiler available for the ECS at the time of the work was able to optimise code efficiently. The use of Occam for the slave nodes should have given near maximal performance for these processors - as Occam was developed at the same time as the transputer. However, the Fortran-77 compiler may not have been able to produce code as efficient as Occam compiled with the Occam compiler.

The implementation of the GROMOS forcefield was not completed. Given time the Occam code would have been converted to Fortran-77, and the Occam communication harness replaced by CS-Tools communications. Both of these changes would have a big impact on program performance - communications via CS-Tools is reported to be a factor of 4 *slower* than via Occam. Performance gains may have been gained by low level programming of the transputer using the assembly language, GUY. The advantage of conversion of the code to Fortran-77/CS-Tools would be its portability. In particular the code could be easily transferred to newer Meiko machines which use vector processing units to increase the performance of the transputer nodes. However, increasing the

computing performance of individual nodes is unlikely to increase program performance if the speed of communications remains the limiting factor in program execution time. This would seem to be the case for the present implementation. Improvements to this would include reducing the amount of data communicated per atom but increasing the amount of data sent at a time (by combining the information for several atoms). Work by others indicates that a minimum packet size for transfer under CS-Tools is 400 bytes (Fincham and Mitchell, 1990).

Chapter 6

Discussion

6.1 Nephropathy

The biochemical work to date suggests that the hydrocarbon induced nephropathy seen in male rats is not a threat to humans, or indeed any other animal. In particular the work of Lehman-McKeeman *et al.*, comparing the detailed differences between the rat and mouse, has shown that the mouse is not susceptable to the same nephropathy even though a highly homologous protein (MUP) is produced (Lehman-McKeenan *et al.*, 1992). Two reasons for this observation have been put forward:

1. MUP is not reabsorbed by the mouse kidney

2. MUP is not capable of binding the nephrotoxic hydrocarbons

The first reason probably stems from differences in function of a2u and MUP in rats and mice respectively. The non-reabsorption of MUP in mice suggests that the excretion of the protein from the body is of prime importance, thus reinforcing the theory that MUP is involved in pheromonal marking of territory or attracting the opposite sex. The reabsorption of a2u in male rats suggests that some internal function is also fulfilled. The sex-related synthesis of a2u implies that any internal function carried out is also sex-related. This has yet to demonstrated, to date only an involvement in fatty-acid transport has been suggested (Kimura *et al.*, 1991).

The second reason must be due to differences in the structure of a2u and MUP. The work of Lehman-McKeeman *et al.* demonstrated that both rats and mice metabolise *d*-limonene to the nephropathic product *d*-limonene-1,2-oxide. They then showed that this metabolite was not bound by MUP either *in vivo* or *in vitro*. There must be some structural differences between a2u and MUP which accounts for this difference in binding. Comparison of the amino acid sequences of a2u and MUP may point to regions of variation which may be where critical differences lie. It must be remembered that the sequences of several different isoforms of MUP have been determined, only some of which are synthesised in the liver (Shanan *et al.*, 1987a and 1987b). It is likely that only isoforms MUPI, MUPII, and MUPIII are relevant to the mouse kidney.

6.1.1 Hydrocarbon Ligand Binding

The hydrocarbon ligands which bind to a2u *in vitro* have been listed in chapter 2 (table 2–5). They are chemically diverse, but all share a common feature - a negatively charged group. The molecules may be generally described as being hydrophobic with an electronegative group. It is interesting to note that not all ligands are active in causing hyaline droplet accumulation (Borghoff *et al.*, 1991). The structures of the more active ligands are presumably complementary to the binding site in a2u. A qualitative structure function relationship for the most active ligands has been determined (Borghoff *et al.*, 1991). Their work is in agreement with work carried out here. The structures of *d*-limonene-1,2-oxide (DLO), 2,4,4-trimethylpentan-1-ol (244T1), 2,4,4-trimethylpentan-2-ol (244T2), 2,2,4-trimethylpentan-1-ol (224T1), α-tetralone, isophorone, retinol, and α-tetralol were superimposed using the FIT option in SYBYL (version 5.4). The electro-negative oxygen atoms from each compound can be placed in the same region of space while maximising the overlap of the hydrophobic portions of the ligands (figure 6–1). The three most potent nephrotoxic ligands; DLO, 244T1, and 244T2, occupy a small steric volume (figure 6–2). The other, less active ligands have groups which lie outside the volume described by the three most active compounds. These groups decrease the energy of interaction with the

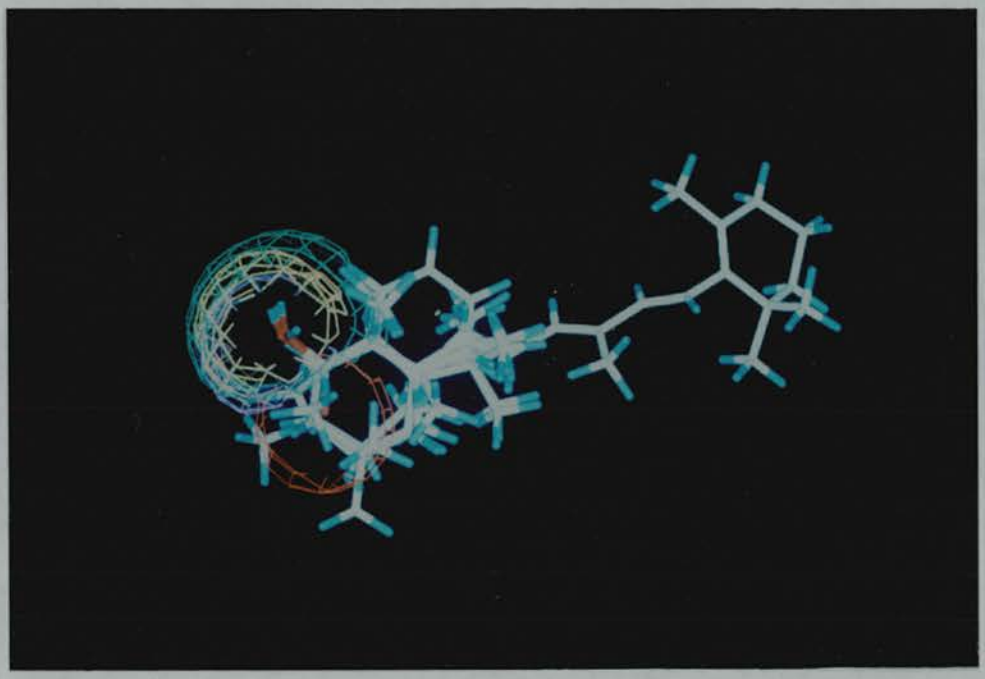


Figure 6–1: Hydrocarbons shown to bind to a2u *in vitro* superimposed.

protein, leading to a decreased binding affinity. That such a broad range of compounds bind, albeit with varying affinities, suggests that the binding site of a2u is variable. This could be because the binding site can change conformation to accommodate different ligands. Alternatively, there are many different species of a2u all with slightly different binding sites. The latter is certainly true - a least six different isoforms of a2u are identified by isoelectric focusing (chapter 3). However, it is unclear as to whether all of these isoforms are active in ligand binding. Biochemical determination of the amount of hydrocarbon bound to a2u recovered from dosed male rats usually only gives a figure of 30% or less. This may be because only one or two isoforms are capable of binding the ligand, or alternatively, the other 70% of the protein may have a tightly bound natural ligand.

6.1.2 The Binding Site

The ligand binding studies indicate that the binding site in a2u will be hydrophobic with a hydrogen bond donor group available to form a hydrogen bond with the electro-negative oxygen of the ligand. It has been demonstrated

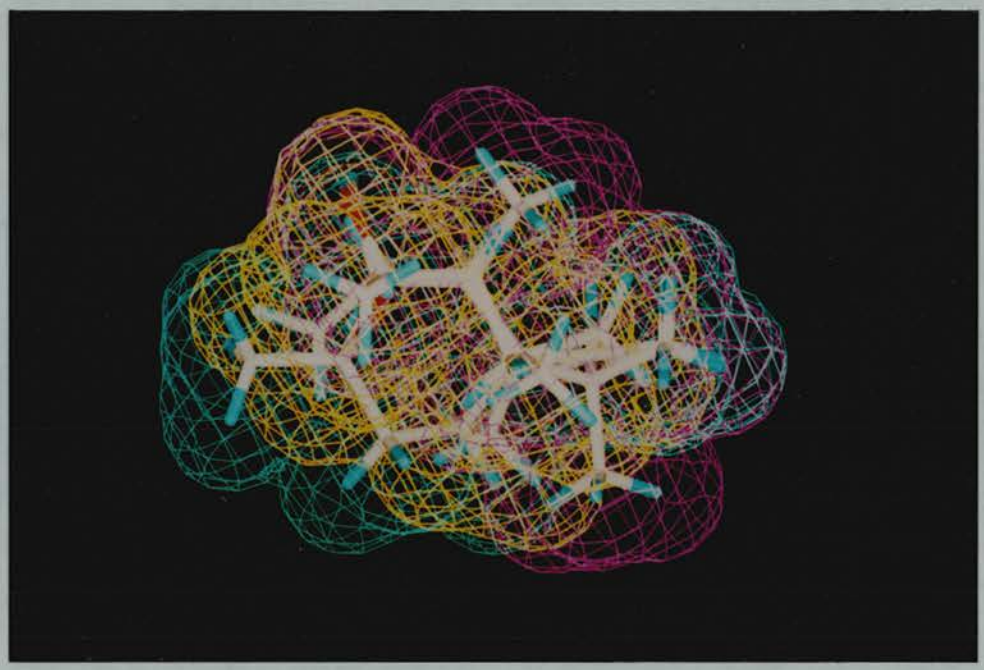


Figure 6–2: High affinity hydrocarbon ligands for a2u superimposed. Molecular volumes shown for T2441 (violet), T2442 (yellow), and DLO (green).

that MUP does not bind *d*-limonene-1,2-oxide *in vitro*. This could be because the binding site is smaller or occluded in some way, or the hydrogen bond donor group is absent. It is also possible that the MUP used in the binding study had a tightly bound native ligand, which the *d*-limonene-1,2-oxide could not displace. Gentle hydrocarbon solvent extraction of the MUP sample before the binding assay might clarify this point. However, it is assumed here that the difference in binding is due to some structural difference between a2u and MUP. Comparison of the sequences of a2u, MUPI and MUPII show some differences (figure 6–3), MUPIII was not included because the sequence is incomplete. There are more changes, mainly conservative, in the first half of the sequences. This agrres with the observation that structural variation in the lipocalycin structures determined to date is in the first β-sheet.

This alignment shows that the distribution of polar and non-polar residues is retained through the three sequences. There are few cases of a hydrogen bond donor in a2u being substituted for a non-hydrogen bond donor in MUP. It is unlikely that a charged residue (e.g. arginine, aspartate, lysine, or glutamate) will be situated in the hydrophobic binding pocket unless forming a salt bridge

```
> mupi EEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQIHVLENSLVLKFHTVR
>mupii EEASSTGRNFNVEKINGEWHTIILASDKREKIEDNGNFRLFLEQIHVLEKSLVLKFHTVR
>  a2u EEASSTRGNLDVAKLNGDWFSIVVASNKREKIEENGSMRVFMQHIDVLENSLGFKFRIKE
      ****** *..*.*.**.* .*..**.******.**..*.*...*.***.** .**..

> mupi DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLY
>mupii DEECSELSMVADKTEKAGEYSVTYDGFNTFTIPKTDYDNFLMAHLINEKDGETFQLMGLY
>  a2u NGECRELYLVAYKTPEDGEYFVEYDGGNTFTILKTDYDRYVMFHLINFKNGETFQLMVLY
      ..**.** .** ** ..*** *.*** ***** *****...* **** *.******* **

> mupi GREPDLSSDIKERFAQLCEKHGILRENIIDLSNANRCLQARE
>mupii GREPDLSSDIKERFAKLCEEHGILRENIIDLSNANRCLQARE
> a2u  GRTKDLSSDIKEKFAKLCEAHGITRDNIIDLTKTDRCLQARG
      **. ********.**.*** *** *.*****....******.
```

Figure 6–3: Sequence alignment of a2u, MUPI, and MUPII, aligned with CLUSTAL, using default parameters.

with another charged residue or a solvent molecule. It is more likely that an aliphatic or aromatic side chain with a hydroxyl group (e.g. serine, threonine, and tyrosine) could exist in an hydrophobic environment and still function as a hydrogen bond donor. There is a change from a tyrosine at 100 in a2u to a phenylalanine in MUPI and MUPII. There is also a change from a threonine at 144 and 154 in a2u to a leucine and alanine respectively in both MUP sequences. These observations still are not enough to determine the differences in ligand binding. Complete understanding can only come from a comparison of the structures of a2u and MUP. Fortunately the coordinates of MUP were made available and the structure of a2u modelled from that of MUP (chapter 4).

The superimposed structures of MUP and a2umup were analysed using interactive graphics. The sequence changes highlighted above were checked to determine where they lay with respect to the suggested ligand binding site - the calyx of the barrel by analogy with other lipocalycin structures. None of the sequence changes noted above lies within the hydrophobic calyx. Both threonine 144 and 154 are on the surface of the protein - exposed to solvent, while tyrosine 100 is at the helix/sheet interface. The differences between a2u and MUP which effect hydrophobic ligand binding may be subtle. Comparison of the calyx shows that the pocket is smaller in a2u (figure 6–4). This is because 2 more phenylalanine rings are present in the calyx (Leu54 $\Rightarrow$ Phe54, and Ala103 $\Rightarrow$

Phe103). This change in the calyx causes the ligand binding site to be shifted further towards the entrance in a2u. This movement may allow the ligands to interact with a hydrogen bonding donor, which they otherwise are unable to do in MUP. That a2u can bind a long hydrophobic molecule such as retinol *in vitro* suggests that the interaction between ligand and protein is extended and hydrophobic. Although the low binding affinity for retinol implies that the interactions may not all be favourable. It is possible that the site of hydrocarbon ligand binding is not the same as the site of native ligands (shown to be the hydrophobic calyx in the crystal structure of MUP). Another possible binding site is the interface between the β-sheet and α-helix. The structure of the trigonal crystal form of β-lactoglobulin suggests that this region is capable of binding retinol (Monaco *et al.*, 1987). However, mutation of relevant residues of BLG implies that the major binding site is the calyx. It is difficult to see how a ligand could bind in a tightly packed region such as the sheet/helix interface. It is always possible that this region becomes more accessible, as the protein undergoes a breathing motion, but this cannot easily be detected in the crystal structure. For ligand binding to occur in the calyx there must be some change in the position of the loop between strands E and F. The sidechain of Tyr84 lies across the entrance of the calyx blocking the entrance/exit of ligands. The crystal structure of MUP has density for a bound ligand in the calyx, suggesting that the loop is mobile enough to allow ligand entry.

The structure of a2u as modelled from MUP does not indicate where hydrophobic ligand binding occurs. The crystal structure of a2u with a bound ligand, such as TMPOH, would solve this problem. Meanwhile, other studies could help narrow the possible sites down. Binding studies with hydrocarbon ligands, and the pheromone based compounds, believed to be similar to native ligands, could determine whether they share a common binding site. Competitive binding assays between [^{14}C]TMPOH and pheromone-like ligands (chapter 2) would indicate if they compete for the same site. The difference in affinity for *d*-limonene (DL) and *d*-limonene-1,2-oxide is in the order of the energy of a hydrogen bond using,

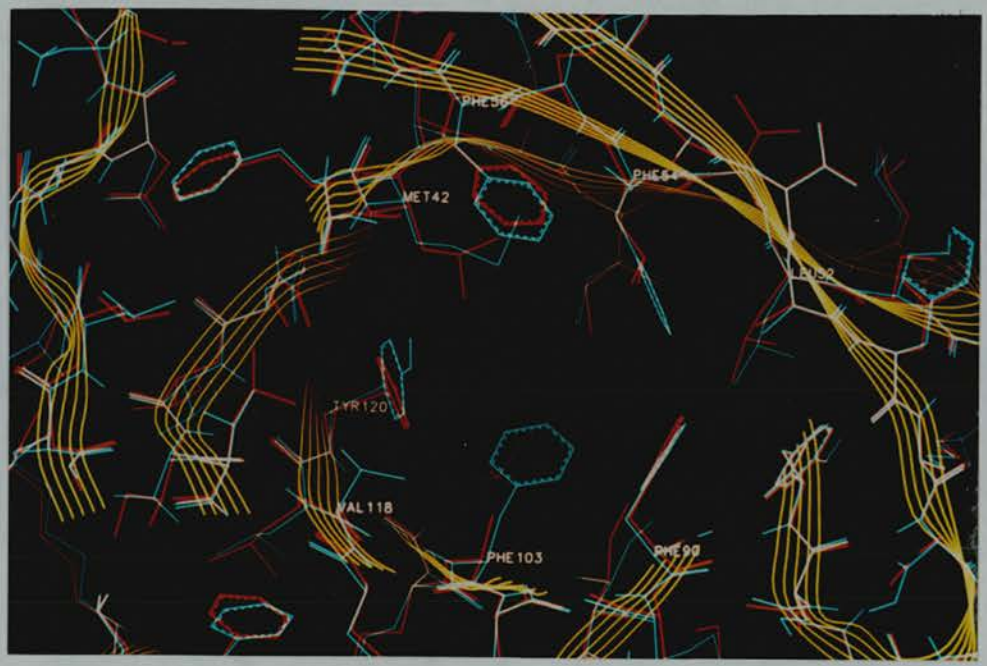


Figure 6–4: Comparison of residues in the hydrophobic calyx of MUP (red) and a2umup (cyan).

$$\Delta G = -RT \ln \frac{K_{i2}}{K_{i1}} \tag{6.1}$$

where K_{i1} is the inhibition constant for DLO, and K_{i2} is that of DL. This gives a difference in binding energy of -14.2 kJ mol^{-1}, or -3.4 kcal mol^{-1}. The energy of hydrogen bonds lie in the range 4-40 kJ mol^{-1}. The only difference between the two molecules is the addition of an oxygen atom.

The high affinity nephropathic ligand, *d*-limonene-1,2-oxide was modelled into the hydrophobic calyx of a2umup using interactive graphics (SYBYL version 5.5). The accessible surface areas of a2umup and DLO were calculated using the program MS (Connolly, 1985). It was possible to dock DLO into the calyx of a2umup with a minimum of overlap of the surface areas (figure 6–5). In addition, the electronegative oxygen atom of DLO lies within 3.0 Å of the hydroxyl group of Tyr120, and the rest of the ligand is surrounded by hydrophobic residues. If DLO is placed in the same place in MUP there is a cavity left by the change of Phe103 to Ala103 (figure 6–6). Possibly this difference is the reason why DLO is unable to bind to MUP - there is insufficient shape complementarity.

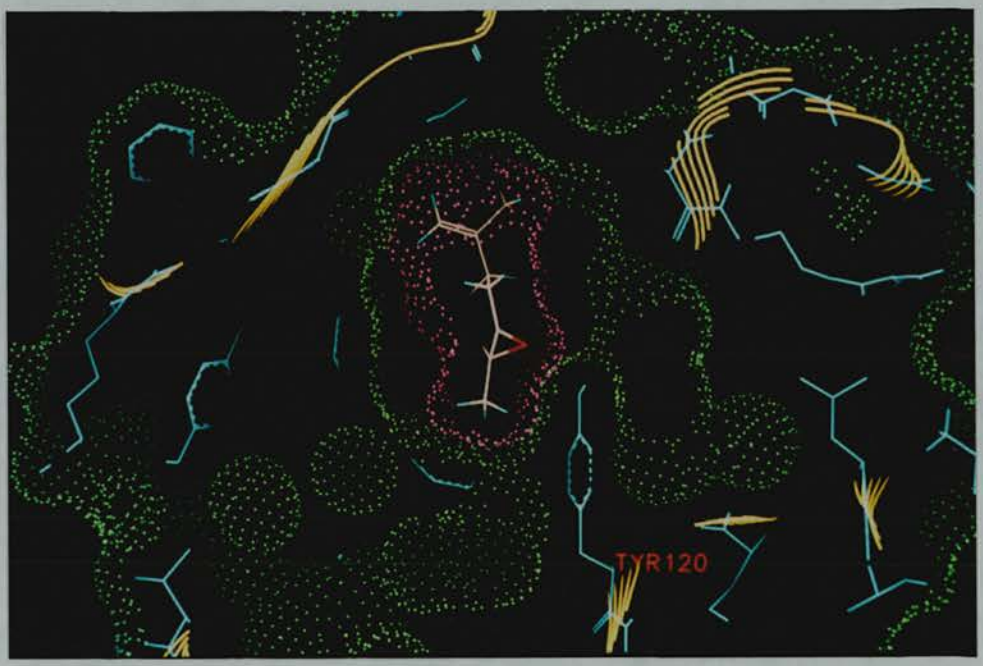


Figure 6–5: Nephrophatic ligand *d*-limonene-1,2-oxide docked into the hydrophobic calyx of a2umup. Connolly surface of DLO (magenta) and a2umup (green).

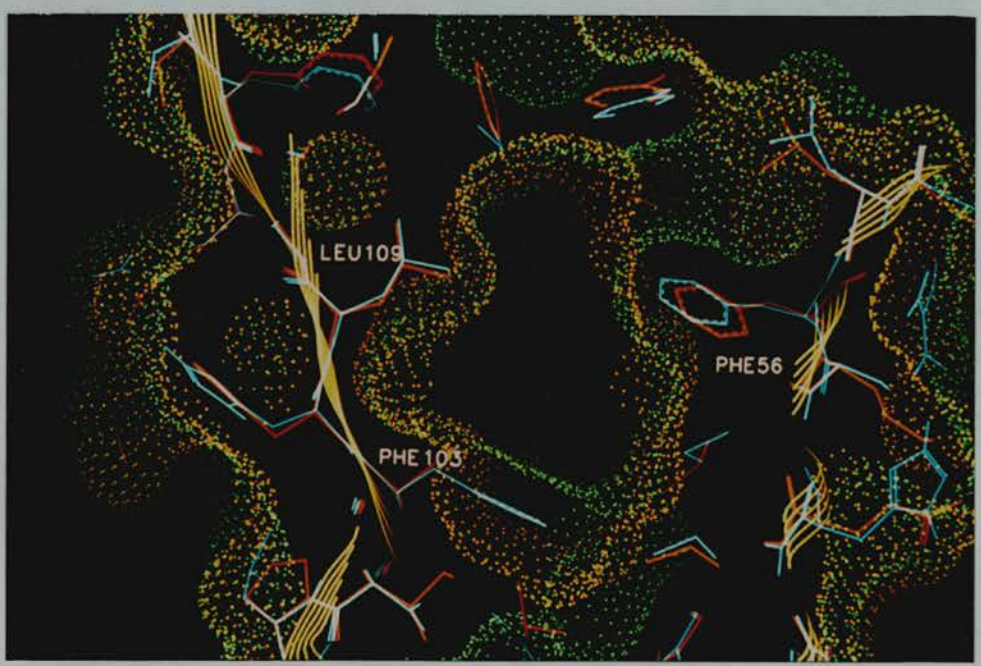


Figure 6–6: Comparison of the solvent accessible surfaces of MUP (red) and a2umup (cyan) in the calyx region. Connolly surface of MUP (orange) and a2umup (green).

6.1.3 Effect of Ligand Binding on Proteolysis

The binding of the nephropathic ligands to a2u causes a decrease in the rate of lysosomal degradation of the complex (Lehman-McKeeman *et al.*, 1990). The same decrease in rate of degradation is also seen *in vitro* with Proteinase K (Charbonneau and Swenberg, 1988), which has a specificity for aromatic and hydrophobic residues. The binding of the ligand must in some way stabilise the protein. Binding could cause a conformational change in the structure of a2u which renders a proteolytic site inaccessible to the active site of the protease. Alternatively, the ligand may itself block the proteolytic site. Work has shown that the ligands themselves are not protease inhibitors (Lehman-McKeeman *et al.*, 1990). Lysosomal enzymes such as Cathepsin D show a specificity for hydrophobic sequences such as: Leu-Tyr, Tyr-Leu, Phe-Phe, Phe-Tyr. Cathepsins C and G also show a specificity for aromatic (Tyr, Phe, and Trp) residues. It is possible that a hydrophobic residue interacts with the hydrophobic ligand such that its conformation blocks, or slows down the action of the lysosomal enzymes. The solvent accessibility of each side chain was calculated by the DSSP program. Very few aromatic/hydrophobic residues show a large accessibility to solvent, as is expected. The residue Phe84 is partially exposed to the solvent, although a shift in the loop away from its position above the entrance to the calyx would expose this residue more. The residues which form the outer-side of strand D are aromatic/hydrophobic and remain exposed. The residues are; Tyr68, Val70, and Tyr72. The full sequence in this region is; L_{67}YLVAY_{72}. Strand D in MUP differs when compared to a2u (figure 6–3). Residue Tyr68 is replaced by a serine residue, while Tyr72 becomes an aspartate. The external surface of strand D in MUP has no exposed aromatic side chains, so is unlikely to be sensitive to lysosomal proteolysis by Cathepsin D. This could be tested by digestion studies *in vitro* with lysosomal enzyme fractions and purified MUP. It is possible that this region in a2u is the first target for cleavage by lysosomal enzymes such as Cathepsin D. Cleavage at this point would open the hydrophobic core of the molecule to further proteolytic attack. The binding of hydrocarbon ligands may keep the core intact after cleavage of strand D. The

docking of DLO into the calyx suggests that favourable van der Waals interactions are formed with hydrophobic side chains, while a relatively strong hydrogen bond is formed with Tyr120. These interactions would serve to keep core side chains grouped together, thus maintaining the structure of the calyx even when exposed to solvent after cleavage of strand D. This observation may explain why some ligands which can be shown to bind to a2u do not cause hyaline droplet accumulation (Borghoff *et al.*, 1991). The ligand may bind in the hydrophobic calyx of the molecule yet not have sufficiently strong interactions with the internal residues, especially in the absence of a hydrogen bond to Try120, to stop the disruption of the calyx structure upon exposure to the solvent. That strand D is a primary target for lysosomal enzymes could perhaps be determined by digestion studies *in vitro* followed by purification of the peptides produced, both in the the presence and absence of hydrophobic ligands such as DLO.

6.1.4 Summary

Analysis of a model of a2u based on the crystallographic structure of MUP can be used to propose a mechanism for the nephropathy observed with some small hydrophobic ligands:

- The ligands bind in the hydrophobic calyx of the molecule. The ligands with highest affinity have shape complementarity with the binding site. Favourable interactions are van der Waals interactions with hydrophobic residues and a hydrogen bond with the hydroxyl group of Tyr120.

- This binding stabilises the core of the molecule. Proteolysis of a2u usually starts at strand D, where Tyr68 and Tyr72 are exposed to the solvent. Normally, cleavage of strand D results in the loss of structure of the core when it is exposed to the solvent, thus allowing further proteolytic breakdown. The presence of the ligand keeps the core intact, for longer if not indefinitely, hence retarding further proteolysis.

- This retarded proteolysis causes accumulation of the protein in the lysosomal vesicle. Over an extended period this accumulation leads to breakdown of lysosomal function. The malfunction of the lysosomes and presence of large aggregates of a2u (hyaline droplets) eventually causes cell death.

- The rapid cell turn-over induced promotes the formation of renal tumors. This is not by direct action by ligand or protein, but rather by the increased chance of mutation incurred when cells rapidly divide and grow.

There are several possible reasons for a lack of hydrocarbon induced nephropathy in mice:

- The hydrophobic calyx of MUP is of a different shape (larger) than that of a2u. Thus the active ligands known for a2u are unlikely to bind as tightly. In addition, it has yet to be shown that MUP does not have a native ligand which is so tightly bound that other ligands cannot bind.

- The suggested proteolytic site, strand D, in a2u does not have the same distribution of amino acids. In particular the exposed tyrosines (68 and 72) are replaced by a serine and aspartate respectively. MUP is therefore either insensitive to lysosomal breakdown, or has different primary proteolytic sites.

- MUP is not reabsorbed by the cells of the proximal tubule. Hence, even if ligands were bound, and some effect on proteolytic stability were produced, lysosomal accumulation of the protein would not occur.

The foregoing is only a hypothesis based on a model of a2u derived from MUP and many assumptions. The crystal structure of a2u with and without a bound ligand, such as DLO, will help test this hypothesis.

6.1.5 Human Nephropathy

It has been suggested that the nephropathy induced in male rats by certain hydrocarbons is not a threat to man (Lehman-McKeeman and Caudill, 1992). The structural analysis above supports this idea, as even a closely related protein such as MUP has sufficiently different features that mice are resistant to the same problem. It seems that the male rat is unlucky in synthesing a2u, which is then reabsorbed and degraded in lysosomes. It is noted however, that in man plasma retinol binding protein is reported to undergo a similar catabolism (Cowan *et al.*, 1990). The levels of RBP are less than a2u in male rats. Is it possible that some ligand of RBP could interfere with its catabolism to such an extent that similar nephropathic effects are produced in man? No such observation has been made, possibly because the binding site of RBP is unable to accommodate ligands other than retinoids. Or, maybe the renal catabolism of RBP is sufficiently different that ligand binding could not effect proteolysis.

The pathology of the human kidney is too diverse to discuss here, but a detailed account is given in Tisher and Brenner (1989). In many cases high doses of small molecules or heavy metals can cause a nephropathy, often generically called chronic tubulointerstitial nephropathies. Cadmium is seen to cause a degeneration in the proximal tubules not dissimilar to hydrocarbon induced nephropathy in male rats. It is suggested that the cadmium forms a complex with metallothionein which is taken up by proximal tubule cells. The complex accumulates in lysosomes leading to cell damage. It is noted that urinary levels of α_1-microglobulin are elevated in kidney disfunction caused by cadmium poisoning (Ekstrom *et al.*, 1975). It is unclear whether the raised level of this protein in the urine is due to the physical damage the kidney incurs, or a an induction of synthesis. It is possible that some specific interaction between the α_1-microglobulin and cadmium takes place, in an analogous manner to the interaction between MUP and Cd^{2+} ions. The A1MG does not seem to accumulate in the proximal tubules and could therefore act as a cadmium clearance vehicle. Alternatively, if the suggestion that A1MG is involved in the immune system (chapter 2) is correct then elevated levels of A1MG may be part

of the bodies reaction to physical kidney damage induced by cadmium poisoning. Bismuth is also reported to form yellowish-brown inclusion bodies in the cytoplasm and nuclei of proximal tubule cells. Lead is also a strong nephropathic metal but its effects are less well defined. Copper-induced kidney damage has been reported in Wilson's disease in man. The effects of copper loading in rats has been investigated (Fuentealba *et al.*, 1989). Both cytoplasmic and nuclear effects were observed. Material accumulated in the lysosomes of the proximal tubules. Two distinct species of lysosome were observed, their origins are thought to be a copper-metallothionein complex, and a copper-a2u complex. The latter is unlikely in man, but the former may play an important part in copper nephropathy.

Human nephropathy is not limited to the action of metals or small molecules. Nephropathy can be caused by the over production of antibodies, in particular IgA (Berger's disease). A condition known as multiple myeloma nephropathy is also reported, and seems similar to Berger's disease. In multiple myeloma nephropathy, proteinuria is observed in which the excreted proteins have been characterised and are called Bence Jones proteins (Sanders and Booker, 1992). These proteins are infact the light chains of monoclonal antibodies which are over-expressed, presumably as a result of the myeloma(s). During the nephropathy these proteins become deposited in the renal tubules in both mice and man. In addition, small crystalline bodies are sometimes formed in the endolysosomal system of the proximal tubules (Solomon *et al.*, 1991). This accumulation of crystalline bodies leads to proximal tubular disfunction.

These results suggest that nephropathies can be induced in humans via a similar mechanism to those induced in male rats by hydrocarbons. A direct analogy cannot be drawn as the proteins involved are either unknown or distinctly different to a2u in male rats. However, a sex-dependent marker of tubular dysfunction has been reported, with a molecular mass similar to that of a2u (Bernard *et al.*, 1989). It is possible that this protein is in fact A1MG, whose concentration is elevated in kidney dysfunction (Ekstrom *et al.*, 1975). The general mechanism of hydrocarbon induced nephropathy does seem applicable to

man. The binding of a ligand to a protein can lead to lysosomal malfunction, protein accumulation and cell damage. Alternatively, some heavy metals may directly effect the function of lysosomes thus producing similar effects. However, at the detailed level, the male rat model is not relevant to man. Proteins similar to a2u exist in man: RBP, A1MG, A1GP, etc. (see chapter 2), but none of these have been shown to have to same biological metabolism or ligand binding capabilities of a2u. In man it is observed that massive exposure to petroleum distillates, may on rare occasions, cause renal failure due to tubular necrosis (Phillips, 1983). However, this condition appears reversible and is not associated with hyaline droplet formation. Hydrocarbons, therefore, do present a risk to the kidney but not of a scale to warrant general concern.

6.2 Structural Studies

Chapter 3 showed some of the difficulties often encountered in X-ray crystallography. The modelling of the structure of a2u was possible (chapter 4) but the reliabilty of such work is still questionable.

6.2.1 Crystallography

Proteins such as a2u and MUP can be obtained in quantities large enough for crystallisation trials. However, the growth of crystals suitable for diffraction studies can be problematic. The problems with a2u are ascribed to the use of impure protein samples. That is to say, the samples were homogeneous with respect to molecular mass but not to charge. The two-dimensional gel electrophoresis showed six charge species of a2u, all of approximately equal proportion. These charge species could have been due to natural isoforms of the protein, by analogy with MUP, or alternatively, due to deamination/oxidation of residues on the surface of the protein during the purification procedure. The crystallisation of *p*-hydroxybenzoate hydroxylase was seen to be irreproducible, even though the crystal structure had been solved by other workers previously

(van der Laan *et al.*, 1989). Biochemical analysis indicated that the purification procedure used introduced oxidation of a free cysteine group. This oxidation allowed multimerisation of the protein through a disulphide bond. The multimers were unable to form stable, ordered crystals. It is unlikely that a2u suffered the same problems but deamination of surface asparagine and glutamine residues could have occurred. Different charged species at equivalent postions on the surface of the protein molecules could make the formation of a stable crystal lattice difficult. The presence of a charged glutamate or aspartate residue instead of a neutral glutamine or asparagine residue could cause electrostatic repulsion between neighbouring molecules - disrupting the crystal lattice. The same argument applies to naturally occuring charge variants.

The purification of the protein to a single charge species may have helped. Although, the growth of crystals of a2u, suitable for structure determination, from an apparently crude purification of a2u has been reported (Böcskei *et al.*, 1991). In the case of MUP, purification to one charge species resulted in the growth of small crystals with a well defined morphology (figure 3–18). The crystallisation procedure used for the determination of the structure of MUP used Cd^{2+} ions to stabilise the crystal lattice, as well as acting as a precipitant. Crystals of MUP were obtained using both ethanol and ammonium sulphate as precipitant (chapter 3). Addition of Cd^{2+} ions to the crystallisation buffer may help produce larger, stable crystals. However, the interaction between MUP and Cd^{2+} would seem to be so specific that the same tetragonal crystal form would be obtained. Alternatively, the addition of ligands to the crystallisation media may help stabilise the structure. The problem with both a2u and MUP is knowing what the natural ligands are. Binding studies have suggested pheromone related compounds do bind , and hydrocarbon ligands from petroleum have also been identified (chapter 2). However, it may not be easy to introduce hydrophobic ligands into the crystallisation media without the presence of quantities of solvents such as ethanol. This may disrupt the crystallisation process from ammonium sulphate, although both a2u and MUP produce crystals from ethanol. Pre-incubation of the protein with ligand may

present a solution to this problem. However, both structures of MUP and a2u are reported to show electron density within the calyx - presumably of a bound native ligand. Whether the protein samples used in chapter 3 also carried bound ligand is not known. Both samples were freeze-dried, a process which may remove a volatile pheromone-like ligand.

The twinning problem observed in crystals of a2u would seem to be insoluble, without some modification to the crystal growth procedure. The level of twinning observed with crystals of a2u grown from ammonium sulphate was such that no reliable interpretation of the data could be made. Twinning is a common problem with protein crystals and the rapid data collection techniques available can be easily used to process twinned data wrongly. The determination of any protein structure should therefore begin with the growth of consistent, well defined crystals. In the case of a2u this may be achieved by further protein purification and changes to the crystallisation conditions.

6.2.2 Molecular Modelling

The modelling of a2u was a modest success. Analysis of the model at the gross secondary structure level showed it to have many of the features of the lipocalycin family. The 8 stranded β-barrel composed of two β-sheets was successfully modelled, with a mainly hydrophobic core. However, analysis of the structure at the level of side chains showed that the model lacked several vital features.

The modelling work highlights several prerequisites for successful modelling. Firstly, and perhaps most importantly, the use of high resolution structures which have been well refined. The use of a incompletely refined structure from low resolution data, such as BLG, shows the subsequent problems in obtaining a native like model. it is important that the source, resolution and refinement of a structure be considered before its use in model building. Standard checks, such as the Ramachandran plot and DSSP algorithm should be applied to assess structure suitability before modelling. When low levels of sequence similarity are

being considered, approximately 25% in the case of the lipocalycins, it is advantageous to use as many related crystal structures in the modelling process as possible.

Assuming a reliable core structure of the model can be constructed the next problem is the completion of loop regions. The loop search algorithm used in the modelling of a2u selects loops on the basis of length and the rms fit of the start and finish α-carbon coordinates with respect to the anchor points on the model. This procedure selects loops which fit anchor points, but takes no account of the surrounding environment. It is still left to the user to decide which loop from a list of many, best fits the structure. This can result in accepting a loop with a poor rms fit to anchor points, low sequence similarity, or a low resolution source structure. It is observed that loops obtained solely by searching for the same fragment length and close geometric fit to the loop base region, will yield realistic backbone coordinates in only two-thirds of test cases (Claessens *et al.*, 1989). The search for loops can be improved by taking account of the residues involved - glycine residues are able to adopt unusual backbone conformations. The selection of appropriate loops from the many initially selected can be automated by energy minimisation (Summers and Karplus, 1990).

When the loop regions have been added, a complete protein main chain is available. The final, and possibly most difficult, task is the placement of side chains onto this backbone. As was seen in the modelling of a2u this process can fail, resulting in a non-native model. Often, modelling programs use the coordinates of the main chain nitrogen, α-carbon, and carbon to determine the orientation and position of the side chain. This is usually by searching a dictionary of side chain rotamers, the one which best matches the main chain atoms is chosen. A new technique has been reported which builds both backbone and side chain atoms onto an α-carbon model simultaneously (Nilges and Brünger, 1991). All other atoms of a residue are assigned to the same position as the relevant α-carbon position. The system is then subjected to a simulated annealing protocol in which the side chains 'grow' out of the α-carbon positions. The energy constants for the covalent energy terms are set initally low, as the

initial conformation is very strained. The usual non-bonded energy terms are altered such that atoms can pass through one another. This technique has been successfully used to model the conformation of the dimerization region of GCN4, a leucine zipper. Other techniques for placement of side chain groups have also been suggested. The placement of sidechains is usually by selection of rotamers from a side-chain rotamer library (Ponders and Richards, 1987). The correct combination of side chain conformations for a molecule may be found, in theory, by selection at testing of all possible rotamer combinations. Obviously this is not computationally possible - there are 2.7×10^{76} combinations for the relatively small protein insulin (Desmet *et al.*, 1992). It is suggested that the number of possible conformations possible can be reduced by application of a dead-end elimination theorem (Desmet *et al.*, 1992). This method rejects side chain rotamers which are absolutely incompatible with the global minimum energy conformation of the molecule. This process reduced the number of possible side chain conformations to 10,800, but still relies upon the accurate placement of backbone atoms. In the modelling of a2u the highly conserved nature of certain side chain conformations could have been maintained by retaining those side chains in the model from the beginning.

In the modelling of a2u by method 4 incorrect loops placements were made. In addition modelling from α-carbon coordinates introduced large errors in the final position of side chains. However, the largest contribution to the inaccuracy of the final model was the misalignment of the sequence of a2u and other lipocalycin sequences. This misalignment is a result of a 4 residue insertion in the loop between strands A and B seen in the crystallographic structure of MUP. This misalignment could not be easily detected in sequence alignments. However, inspection of the final model revealed an arginine residue (Arg44) and a glutamate residue (Glu33) inside the hydrophobic calyx of the molecule. This anomaly suggested that the model was incorrect but did not indicate how it had to be changed. Only when the structure of MUP became available did the solution become apparent. Techniques for assessing the correctness of structures based upon the environment of each amino acid have recently been developed

(Lüthy *et al.*, 1992). This kind of test would have indicated that there was a problem with the misthreading of the strands. It may be possible to correct such mistakes by an extensive search through rethreaded conformations. This process can be automated to some degree, by assuming that the position of strands is correct and only their threading is incorrect (Bowie *et al.*, 1991; Thornton *et al.*, 1991; Jones *et al.*, 1992).

The long loop between strands A and B in MUP is novel compared to the other lipocalycin structures solved to date. This information has been incorporated into the sequence alignment of the lipocalycin structures solved so far (chapter 4). The loop between strands A and B in MUP contains a 5 residue insertion compared to RBP (figure 4–33). The sequence alignment tools commonly available are unable to align MUP and RBP correctly. An assessment of the reliability of different regions of the alignment may have indicated a problem in this loop region (Vingron and Argos, 1990). Secondary structure prediction of a2u shows α-helix for residues 27 to 34, the residues between strands A and B. However, the reliability of secondary structure prediction (60%) did not give enough confidence to incorporate such information in modelling. However, the analysis of model 4 indicated a problem in the region of this loop. Had the structure of MUP not become available model 4 would have been remodelled in this region. As suggested above, this remodelling can be automated by varying only part of the model. Future modelling of other members of the lipocalycin family may be most efficiently carried out using the sequence/fold matching method of Jones *et al.*, 1992. The method uses a dynamic programming algorithm to find the best threading of a sequence through the coordinates of a known protein structure. This involves a process of optimising the fit of a sequence to a specific structural motif. In the case of MUP and RBP the structural motif is the lipocalycin fold, the optimal threading for MUP is with a 5 residue insertion in the loop betwen strands A and B. The method has been successful in identifying structural similarity between C-phycocyanin and sea hare myoglobin (Jones *et al.*, 1992). However, other cases

are not successful although the method may be improved by the inclusion of multiple sequence data and the generation of model folds.

6.3 Parallel Processing

The use of both MIMD and SIMD parallel computers was seen to give performance improvements over conventional serial machines. Some algorithms are more suited than others to both data parallel or MIMD parallelism.

6.3.1 Molecular Dynamics

Implementation of the GROMOS force field on a Transputer based MIMD computer gave a maximum increase in performance of 40-fold compared to a VAX 11/750. This required much of the code to be written from scratch, to suit the programming environment available at the time. The implementation of the program was time consuming and complex which is a general problem with message passing parallel systems - there are many new programming concepts which must be mastered. The debugging of code during development is also often difficult when there are a large number of active processors, since it is very difficult to determine the state of data distributed across several processors. This problem is now less important as better debugging, and performance evaluation tools become available. Even though code development took some time, its subsequent use made it worthwhile. However, the performance obtained is no longer competitive with serial machines available today. High performance Unix workstations are typically capable of operating at 10 to 20 times the speed of a VAX 11/750. The code could have been optimised, as outlined in chapter 5. However, as the number of different parallel platforms increases, portability of code becomes a major issue. It is desirable to write code in standard Fortran-77 or C which can then be easily transferred from machine to machine.

The problem with generalised molceular dynamics on any parallel machine is the need to calculate the interactions between particles separated in space. This

leads to much global communication which is time consuming in any system. The increase in performance of single processors has not been matched by an increase in communications performance. In the parallel implementation of GROMOS the bottle-neck was communication, hence increasing the processing power of each node would be unlikely to provide a dramatic performance increase. Although, increasing the amount of data communicated at each step would have a two-fold benefit. Firstly, the overhead incurred in starting up a data communication (15 μs using Occam, 400 μs using CSTools) is minimsed with respect to the time taken to transmit the data (1 MByte/s using either Occam or CSTools). Secondly, the larger data packet size would allow vector pipelining of the calculations on the processing node, assuming this kind of optimisation is supported. Porting of a Fortran/CSTools version of the algorithm to an i860 based machine would therefore provide some increase in performance.

The main problem with the implementation at present is its inherent lack of scalability. The communication of all data twice around the ring is inefficient, but does allow the use of the SHAKE algorithm to increase the length of the integration time step.

Parallel molecular dynamics algorithms have been implemented by several other groups. The majority of work has focused on Transputer based systems, because of their availability and low inital cost. The basis for much of the work has been the systolic loop algorithm (Ostlund and Whiteside, 1984). The term systolic was used to describe the pulsed nature of data flow through the system. A ring of processors has approximately the same number of resident particles. Data are passed from a processor to its neighbouring processor. The force between the external particles and resident particles is calculated, and the particles passed on to the next processor. A systolic pulse therefore consists of data communication followed by calculation. For n processors, the number of pulses need only be $n/2$ because of the reciprocal nature of the force calculation between particles. The initial implementation of Ostlund and Whiteside was on a custom built system based around Intel microprocessors.

More recently a similar algorithm was implemented on the ECS (Raine *et al*,

1989). This work demonstrated that the number of particles per processor has an effect on efficiency, more that 10 are need per processor to achieve 90% efficiency. The initial implementation was used to study a purely Lennard-Jones system, argon, but later developed for the protein crambin (Raine, 1991). The performance obtained for crambin simulations was not as impressive as expected. This was partly due to load balancing problems across the processors when a cut-off scheme was implemented, and also the large book-keeping overhead incurred in describing the topology of the molecule. The time per integration step for crambin was approximately the same as that observed for MD8 (0.6 seconds). The communications per integration step were kept at a minimum by integrating the equations of motion for each particle on its home processor. This allowed the hydrogen bonds to be constrained with the SHAKE algorithm, permitting a time step of 1 fs. The ability to SHAKE hydrogen bonds required that the protein be defined in terms of groups ensuring that bonds involving hydrogen atoms were not split across processors. The implementation of the bonded force calculation across the processors resulted in the program being specific to crambin, and a maximum of 23 processors.

The Ostlund and Whiteside topology has also been implemented on custom-built Transputer hardware (Heller *et al*, 1990). A similar systolic loop algorithm was used. However, the accumulation of forces used a second communications ring, so that communication of coordinates and forces occurred in parallel. A multiple timestep algorithm was used in the calculation of long-range forces. This can be considered as an extension to the twin-range method used in GROMOS87. The force between particles separated in space varies only slowly compared to particles which are close together. Therefore, it is argued that the forces between distant atoms only need be updated every few steps, while close atoms are updated every step (Grubmüller *et al.*, 1991). A complex non-bonded pair list is constructed for each atom using a set of 8 distance classes. Neighbouring atoms are in the first distance class and the force is updated every time step. Atoms in the following classes are updated every 2, 4, 8, 16, 32, 64, and 128 steps. The force between all atoms is therefore only calculated every 128 steps, at this point

the non-bonded pair list is recalculated. The calculation of non-bonded terms therefore becomes proportional to $N\log(N)$ as opposed to N^2. The combination of this multiple timestep technique and the overlapped communications/calculation results in impressive performance. The code for EGO was implemented on the ECS by H. Heller and L. Clarke. Timings for simulations of crambin with both MD8 and EGO are compared in figure 6–7. Similar simulations with a2u and MUP show that EGO performs between 2 and 3 times faster than MD8. It was not possible to simulate the large MUP plus solvent system used for MD8. Performance is expected to be less impressive for this system because the program simulates solvent molecules explicitly (i.e. all bonded and angle terms for solvent molecules are also calculated).

Another application of parallelism to molecular dynamics has not used the systolic loop method (Schreiber *et al.*, 1992). A transputer system was used with a communications based on a Fibonacci-tree. As with MD8, the PROMD program from GROMOS87 was the initial code source. The current implementation, written in 3L Fortran, calculates the non-bonded interaction between protein-solvent and solvent-solvent in parallel. The test case under study was a 17 amino acid peptide surrounded by 1021 water molecules. Timings for an integration step on 18 transputers are reported to be slightly less than for PROMD running on an IBM 3090. The minimisation of interdependence of communication and computation allowed the code to be readily ported to a workstation environment using the Linda distributed processing system (Scientific Computing Associates, 1992).

All of the implementations discussed have been on MIMD systems, often transputer based. This highlights a problem with generalised molecular dynamics algorithms - they are not easily implemented on SIMD systems. The fundamental problems are the global communications needed to calculate long-range forces, and the need for each processor to execute the same code. The implementation of neighbour lists is therefore problematic. However, if all pair-wise interactions are to be calculated it is possible to use a SIMD algorithm, this has been carried out already (Windemuth *et al.*, 1991). The replicated

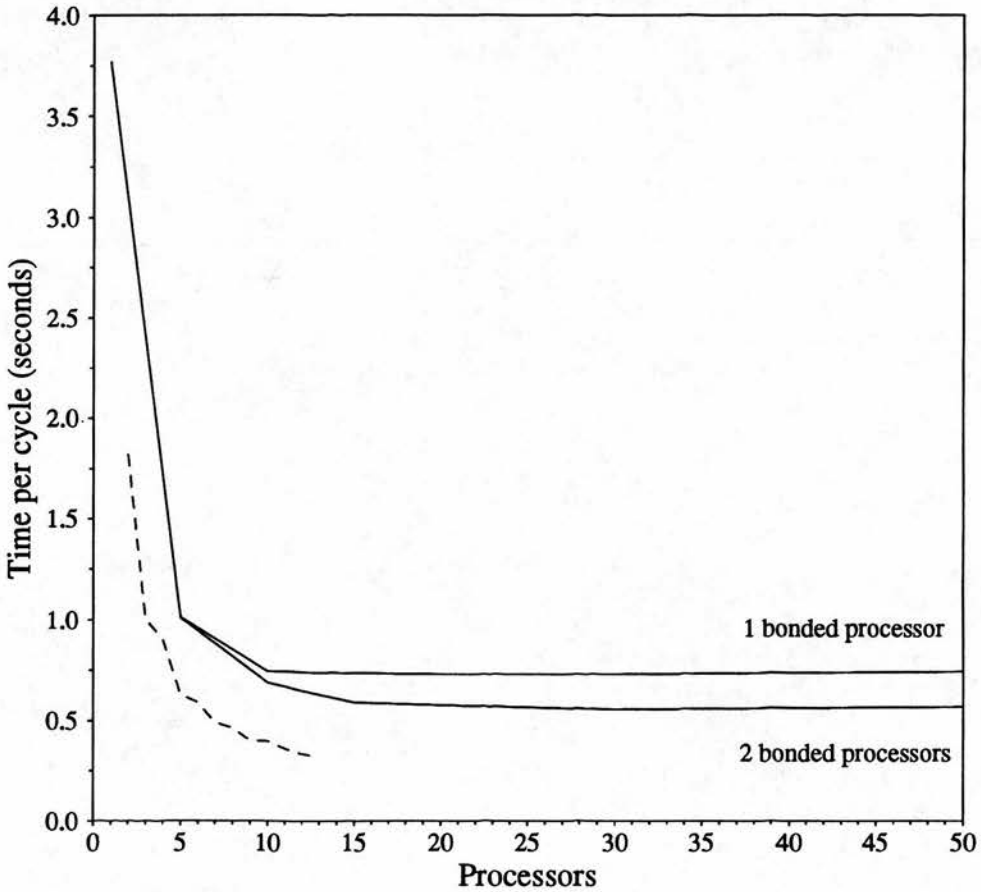


Figure 6–7: Comparison of the time per integration step for crambin using both the programs EGO (dashed line) and MD8 (solid lines).

systolic loop algorithm was used with all N_a atoms per loop and M copies of the loop. The algorithm was implemented in C/Paris (Thinking Machines, 1989) on a 32K CM-2. Pairwise interactions within a loop are calculated by cyclic shift atoms to neighbouring processors, only $N_a/2$ shifts being needed to calculate all forces. For each replicated loop the atoms are preshifted such that each loop starts with a different pairing of atoms. Therefore, with 3 loops the complete set of interactions can be calculated in $N_a/6$ steps. The algorithm is efficiently coded in a data parallel manner using spreads, cshifts, and sums. Only NEWS communication is required for this kind of implementation. Performance results suggest a loop replication of $N_a/10$ is the maximum usable. The reported implementation of this algorithm used C/Paris which although relatively low-level would not be able to make full use of the FPA units. Conversion of the algorithm to slicewise CM-Fortran would be possible but is not clear how this would be any improvement over a single systolic loop. There are only a maximum of 2048 FPAs in a CM-200 giving 8192 vector operations at one time. Implementation of a single systolic loop under slicewise CM-Fortran would be the first step, followed by tests to determine the gains from replication. The calculation of all pairwise interactions on a SIMD machine is possible and rapid. However, the calculation of the bonded terms is less easy to decompose in such a regular manner. The problem is that bonded terms involve local, but irregular communication. Therefore, to calculate the angle energy terms information from neighbouring processors must be gathered in an irregular way resulting in send/get operations. In addition, some atoms may be involved in more than one dihedral bond resulting in a need for data duplication. It is possible that the atoms could be sorted in such a way that only regular local communication is required to calculate covalent terms. This is unlikely given the branched nature of polypeptide chains and cannot be assumed to be possible in all cases. A rather easier option is to carry out the calculations on the front-end and just use the CM for calculation of the non-bonded terms.

The implementation of PROMDL presented in chapter 5 lies somewhere between the systolic loop and pipelining algorithms. Both have proved successful, the

latter being purpose-built for gravitational many-body calculations (Sugimoto *et al.*, 1990). Although the hybrid algorithm did produce a significant performance improvement its lack of scalability renders its long-term useful doubtful. It would seem that MIMD, or very specialised pipelined hardware, is best suited to molecular dynamics calculations. However, improvements in the scope of dynamics calculations seem likely to come from algorithmic rather than computational developments. The use of multiple timesteps (Grübmuller *et al*, 1991) and multipole techniques (Greengard and Gropp, 1990) may provide a 100-fold increase in the speed of molecular dynamics calculations.

6.3.2 Crystallographic Refinement

It was possible to increase the performance of the least-squares refinement program PROLSQ. However, the usefulness of the program once parallelised is questionable. The refinement is not usually applied in a cyclic manner for many steps before manual intervention is required. The bottleneck in the refinement therefore remains that of interactive model building. The program PROLSQ has been used as a crystallographic benchmark program to assess the performance of different computers (Bourne and Hendrickson, 1990). In this context parallelisation of the code is worthwhile. However, much of the code remains difficult to parallelise on a SIMD machine as global communication is required.

Heteroarchitecture MD refinement

The conventional refinement outlined previously is time-consuming, because the limited radius of convergence of least squares algorithms necessitates periodic examination of electron density maps followed by manual rebuilding of the model. The radius of convergence is given theoretically by $\frac{d}{4}$, where d is the highest resolution reflection included in the least squares refinement (Jack and Levitt, 1978). In general, stereochemically restrained least squares refinement does not correct residues that are misplaced by more than 1 Å. In addition, the

refinement procedure is easily trapped in local minima so that human intervention is required to flip a peptide bond for example.

The technique of simulated annealing (Kirkpatrick *et al.*, 1983) has been used to overcome both local minima and increase the radius of convergence of refinement (Brünger *et al.*, 1987). Molecular dynamics was incorporated into crystallographic refinement by addition of an artificial potential energy term based on the agreement between observed and calculated structure factors:

$$E_{sf} = S \sum_{hkl} [|F_{obs}(hkl)| - |F_{calc}(hkl)|]^2 \qquad (6.2)$$

where S is a scale factor chosen such that the gradient of E_{sf} was comparable in magnitude to the gradient of the empirical potential energy normally used in molecular dynamics (see chapter 4). This technique has been called MD-refinement and has become the preferred method for initial crystallographic refinement, with about 50% of macromolecular structures determined by either crystallography or NMR being refined in this way in the last two years (A. T. Brünger, personal communication).

The technique requires that the crystallographic potential energy term is calculated, usually using the Fast Fourier Transform (FFT) method (Cooley and Tukey, 1965; Ten Eyck, 1977; Agarwal, 1978). The derivative of the difference between observed and calculated structure factors with respect to the atomic coordinates can be readily calculated by convolution of the difference map with the model electron density (Agarwal, 1978). This derivative can be converted to a gradient and hence a force on each atom by use of an appropriate scaling factor. The simulated annealing requires molecular dynamics simulation at elevated temperature (2000 K and above). Typically the system is simulated for at least 1 ps with 1 fs timesteps then slowly cooled to 310 K. Energy minimisation is required prior to and following this procedure. The simulated annealing method allows the system to overcome local energy minima, the likelihood of this being proportional to the "temperature" of the system. In addition, provided the scale factor for E_{sf} is chosen correctly, the model moves

through configuration space in a stereochemically sensible way. This minimises the need for manual intervention in the refinement process.

The technique of MD-refinement provides a much improved radius of convergence (5 Å has been reported (Brünger *et al.*, 1987)) and also minimises manual intervention. However, the process is CPU intensive. Refinement of crambin, 46 amino acids, for a total time of 10 ps, required 1 hour CPU time on a CRAY 1 (Brünger *et al.*, 1987), foot and mouth disease virus required 190 CPU hours on a Convex C210 (Taylor, 1989). Refinement of BLG, for 2.5 ps, required approximately 6 hours CPU time on an ESV20. Structure factor calculations account for about half this time. It is assumed that molecular dynamics calculations account for the other half.

Initial work was undertaken to explore the possiblity of using two different parallel machines in a co-operative manner to carry out MD refinement efficiently. The application of a MIMD platform to molecular dynamics had already been demonstrated (chapter 5). However, the efficient use of MIMD systems for FFT calculations requires some programming effort. A topology well suited to FFTs is a hypercube (Chamberlain, 1988), and is thus quite distinct from the ring topology used in many parallel molecular dynamics algorithms. SIMD machines on the other hand, can be very efficient for the FFT algorithm. The Connection Machine's underlying architecture is a hypercube. The AMT DAP also provides an efficient implementation of the FFT algorithm. Both the CM-200 and DAP can calculate radix-2 FFTs using subroutine library calls. FFTs can be multi-dimensional, and any power of 2 in any dimension for CM-200 library routines (Thinking Machines Corporation, 1992). The AMT DAP was more limited, only allowing calculations in 1 or 2 dimensions with 4096, or 64 x 64 elements respectively (Active Memory Technology, 1989).

Thus, by connecting the ECS and AMT DAP with a high speed parallel interface (HiPPI - high performance parallel interface) most efficient use could be made of the two types of parallelism. Molecular dynamics calculations would be carried out on the ECS. At the appropriate time coordinates would be passed to the DAP via the HiPPI interface. These coordinates would then be used to

calculated structure factors using a parallel FFT. The difference between the calculated and observed structure factors would then be used to calculate forces on each atom. This 'X-ray' energy term would then be passed back to the ECS and included in the total energy. This would be repeated until the refinement was completed. The MD refinement code based around GROMOS87 (MDXREF) was used as a basis for development. This uses the Agarwal FFT based method to calculate atomic derivatives with respect to the structure factors (Agarwal, 1978). This requires that the model electron density be sampled on a regular grid, which is then Fourier transformed to give structure factors. A core code was constructed to take a set of coordinates and calculate structure factors on a 64 x 64 x 64 grid. The time taken for the complete complex $\Rightarrow$ complex three-dimensional FFT was approximately 2 seconds. However, the calculation of the model electron density was very time consuming, taking approximately 4 minutes. This was in part due to the lack of floating point hardware on the DAP which requires all real arithmetic to be carried out in software, and also the need to use costly exponential functions rather than look-up tables to calculate the electron density. It is not possible to do parallel indexing of parallel arrays in DAP Fortran. The same algorithm was implemented on the CM-200 and was considerably faster, approximately 9 seconds for model electron density calculation and 0.2 seconds for the FFT calculation. This still does not provide a significant improvement over conventional serial machines, however, the speed of the electron density calculation could be increased by the use of look-up tables on the CM-200 (Thinking Machines Corporation, 1991b). Provided the structure factor calculation was improved the use of both parallel machines together for MD refinement would be feasible.

The method outlined above uses two parallel computers in serial - they each have to wait for information from the other until they can carry out their calculations. The MD refinement technique has been extended to include time averaging of structure factors (Gros *et al.*, 1990). The 'calculated' structure factors used to derive the X-ray forces are not based solely on the structure at that instant, as

they usually are in classical MD refinement. Instead, a memory function includes information from previous structure factors, thus

$$E = E_{phys} + \sum_{s}(|F_o(\mathbf{s})| - k|\langle F_c(\mathbf{s})\rangle|)^2 \tag{6.3}$$

where

$$\langle \mathbf{F}_c(\mathbf{s})\rangle_{t'} = \frac{1}{\tau_\chi(1 - e^{-t'/\tau_\chi})} \int_0^{t'} e^{-(t'-t)/\tau_\chi} \mathbf{F}_c^t(\mathbf{s}) dt \tag{6.4}$$

An ensemble of structures is calculated during the crystallographically restrained MD simulation. In the test case used, bovine phospholipase A_2, the R-factor drops to 10%, some 7% lower than the classical MD refinement. This is presumably because the time-averaged structure factors model reality more closely than one individual conformation. This is expected as X-ray diffraction data must include an element of dynamic motion, because data are collected over a relatively long time period. The interesting feature of this time-averaged MD refinement is that the molecular dynamics and structure factor calculations can be decoupled to some degree. That is to say, they no longer are serial - structure factors can be calculated while the simulation is continued. This technique is ideal for parallelism because this decoupling can be carried out, and because the simulation length required is in the order of 50 ps some way of speed-up the calculation is required. The same heteroarchitecture system could be used, but the event sequence would be changed. The simulation would generate coordinates which would be sent to the DAP every so often. Once received the structure factors would be calculated, added to the time-averaged set and derivatives calculated. The forces would then be passed back to the ECS to be incorporated into the total force. During the structure factor calculation the dynamics simulation would proceed generating more conformations - the system would be truely parallel. The balance of work between the two machines would need to be controlled so that the dynamics simulation did not get ahead of the structure factor calculation by a large amount.

6.3.3 Future Applications of Parallelism to Protein Structure Determination

The applicability of SIMD parallelism to the brute force solution of the translational parameters in molecular replacement has been demonstrated (chapter 5). Many of the time consuming processes in protein structure determination are open to parallelism. In general the MIMD programming paradigm is favoured. However, as compiler technology advances it is likely that both MIMD and SIMD programming styles will be implemented on an underlying MIMD architecture. The application of massively parallel processing to structural problems will provide the chance to explore methods hitherto unused because of the computational time required.

Molecular Replacement

It is assumed that the technique will remain as two distinct three-dimensional searches for the immediate future. The vector space rotational search as used in X-PLOR (Huber, 1985) is difficult to parallelise on SIMD machines. The method requires the comparison of two Patterson maps, one of which is rotated by small increments relative to the other. This involves a non-local communication of data on a SIMD machine, and is therefore inefficient. MIMD hardware is therefore preferred for the vector space rotation search. The obvious, and easiest, decomposition of the problem onto MIMD hardware is to divide the rotation search into several smaller sub-searches. The number of sub-searches would be the same as the number of processors available. This requires that every processor has a copy of both Patterson maps - this is not likely to be a major problem unless a very large unit cell is being studied. The peaks from each sub-search would be combined and processed by the controlling node. The advantages of this system are scalability and independence of processor numbers.

The phase translation function, as implemented in BRUTE and X-PLOR, is highly parallelisable on SIMD systems (chapter 5). Other translation searches are often based around an FFT and therefore are probably equally applicable to

either MIMD or SIMD machines, but probably most easily coded for on the latter. The phase translation method implemented on the CM-200 could also be applied to MIMD architectures. By analogy with the rotation search, the translation search could be divided into several sub-searches. The sub-grids covered would then be recombined and analysed by the controlling node.

Direct Phase Determination

The limiting factor in crystallographic structure determinations is often the generation of initial phases. The molecular replacement technique attempts to solve this problem using information from similar structures already solved. In general however, the isomorphous heavy atom replacement technique is still used. The ideal solution would be direct phasing *ab initio* based solely on the experimental data collected. Direct phase determination using a combination of maximum entropy and maximum likelihood methods has been reported for a small protein (Gilmore *et al.*, 1991). A maximum entropy method has been successful in the phasing of recombinant bovine chymosin (Sjölin *et al.*, 1991). The later method is inherently parallelisable using the MIMD programming paradigm. In general, direct phase determination is a problem of optimisation which may be solved by extensive testing of different phase combinations. MIMD parallelism is readily applied to all search methods which treat search paths independently.

Protein Structure Refinement

The refinement of both crystallographic and NMR structures requires extended sampling of conformational space. The determination of NMR structures usually proceeds by simulated annealing of random conformations restrained by the through-space distance restraints determined experimentally. Consistent convergence to a common structure from several different starting point is used as a criterion for a valid structure. The greater the number of independent SA runs that can be carried out, the greater the chance of finding the ensemble of

conformations that best describes the structure. The SA runs are all completely independent and therefore can be farmed across a number of processors, m SA runs can be carried in parallel on n processors with m/n runs per processor. The same procedure can be used for crystallographic refinement, using slightly different starting conformations for each SA refinement. This method can only realistic be implemented on MIMD machines.

Molecular Dynamics

In general the searching of conformation space through molecular dynamics does not require the trajectory to be described in detail. What is generally of importance is the result of the simulation, or some general property of the protein throughout the simulation. Hence refinements of the kind outlined above can be carried out as several independent tasks rather than one long one. However, in some cases a long single simulation is required. The determination of free energy differences by cycle perturbation methods usually require extended simulation often in a highly solvated system (McCammon and Harvey, 1987; Wong and McCammon, 1986; Kollman *et al.*, 1987). The systolic loop methods described above provide a way for parallelising the calculation of a single trajectory on MIMD machines. The limitation to this method is the ratio of particles to processors. Studies so far indicate that at least 10 atoms per processor are required for efficiency. When the number of processors equals the number of atoms no further parallelism can be achieved. When the number of processors greatly exceeds the number of atoms the replicated systolic loop algorithm implemented on the CM-2 may be the preferred method of parallelism.

Monte Carlo Methods

Molecular dynamics simulations do not usually sample configuration space completely enough to describe all conformations accessible to a protein. This can lead to misleading results, especially when considering some microscopic property such as the conformation of a side-chain. Increasing the length of the

simulation is not seen to remidy this problem. The application of Monte Carlo techniques (Metropolis *et al.*, 1953) may help explore configuration space more fully. The Monte Carlo algorithm is inherently parallelisable on MIMD systems. Different configurations are generated on different processors. Simulation of each molecule may be used to explore local configuration space. Some criteria, such as lowest total energy, may then be used to select the starting point for the next set of structures. Some communication is required to broadcast this new starting configuration to all processors, but this is not a problem provided it remains small in comparison to the time taken for the simulation. This method may provide a more rigorous way of exploring conformation space in the refinement of protein structures.

Modelling and Folding

With sufficient computing power, provided by massive parallelism, in the future it may be possible to model structures *ab initio* without experimental input other than the protein sequence. Finding the structure of a protein computationally would seem to be a problem of extensive searching of configuration space. Massive parallelism will certainly increase the amount of searching that can be performed, however, it is still not clear that such searches can find the global minimum or indeed that this minimum need necessarily be the native protein structure. The inclusion of other information in the modelling of structures *ab initio* may serve to both speedup the process and direct the model towards a conformationally sensible minimum. The application of parallelism to the process of homology modelling may produce more reliable results. The threading procedure outlined above is implicitly parallelisable, as is the testing of different loop structures in loop searching algorithms.

6.3.4 Conclusion

Clearly, parallel processing has several worthwhile applications to protein structure determination. In general, the use of MIMD hardware and message

passing software is favoured for structural problems, although in some cases the SIMD paradigm can be highly efficient. The use of custom built hardware, although often producing significant speedup, is less and less desirable as parallel hardware converges to a relatively similar design. Effort needs to be placed in the design of general parallel algorithms which can migrate from machine to machine. The cost efficiency of massively parallel processing compared to conventional supercomputers cannot be overlooked.

6.4 General Conclusion

This work suggests that molecular modelling still cannot rival the direct techniques of X-ray crystallography or NMR. Also, correct interpretation of a protein's function requires knowledge of the native structure as well as biochemical data. The study of protein-ligand interactions is particularly sensitive to errors in protein structure. Such errors can be minimised during modelling only by careful consideration of all the information available. A modelled structure should be corroborated by determination of the crystallographic structure if at all possible. Hopefully, the future application of parallel computing to molecular modelling will make the process more accurate. The more immediate use of parallelism in protein structure determination will yield more reliable results in a shorter time.

Chapter 7

References

- Active Memory Technology, Transform Library. Active Memory Technology Ltd., Reading (1989)

- Agarwal R. C., A New Least-Squares Refinement Technique Based on the Fast Fourier Transform Algorithm. Acta Cryst. A34, 791-809 (1978)

- Åkerström B., and Lögdberg L., An intriguing member of the lipocalin protein family: α_1-microglobulin. Trends in Biochem. Sci. 15, 240-243 (1990)

- Åkerström B., Immunological analysis of α_1-microglobulin in different mammalian and chicken serum. J. Biol. Chem. 260, 4839-4844 (1985)

- Alden C. L., Kanerva R. L., Ridder G. and Stone L. C., The Pathogenesis of the Nephrotoxicity of Volatile Hydrocarbons in the Male Rat. In Proceedings of the Workshop on the Kidney Effects of Hydrocarbons, Hemstreet G. P., Thorpe J. J., and Kane M. L., eds. Pg 154-170. American Petroleum Institute, Washington, D.C. (1983)

- Ali S., and Clark A. J., Characterization of the gene encoding ovine beta-lactoglobulin. Similarity to the genes for retinol binding protein and other secretory proteins. J. Mol. Biol. 199, 415-426 (1988)

- Andrews A. T., Electrophoresis: theory, techniques, and biochemical and clinical applications (2nd edition). Oxford Clarendon Press (1986)

- Anfinsen C. B., Principles that govern the folding of protein chains. Science 181, 223-230, (1973)

- Antakly T., Lynch K. R., Nakhasi H. L., and Feigelson P., Cellular dynamics of the hormonal and developmental induction of hepatic α_{2u}-globulin as demonstrated by immunocytochemistry and specific mRNA monitoring. Amer. J. Anat. 165, 211-244 (1982)

- Argos P., A sensitive procedure to compare amino acid sequences. J. Mol. Biol. 193, 385-396 (1987)

- Arndt U. W., and Wonacott A. J., (eds.) The rotation method in crystallography. North-Holland, New York (1977)

- Bacchini A., Gaetani E., and Cavaggioni A., Pheromone binding proteins of the mouse, *Mus musculus*. Experientia 48, 419-421 (1992)

- Baumann G., Frömmel C., and Sander C., Polarity as a criterion in protein design. Protein Engineering 2, 329-334 (1989)

- Bedard P. A., Balk S. D., Gunter H. S., Morisi A., and Erikson R. L., Repression of quiescence-specific polypeptides in chicken heart mesenchymal cells transformed by Rous Sarcoma virus. Mol. Cell. Biol. 7, 1450-1458 (1987)

- Bedard P. A., Yannoni Y., Simmons D. L., and Erikson R. L., Rapid repression of quiescence-specific gene expression by epidermal growth factor, insulin, and pp60^{V-src} Mol. Cell. Biol. 9, 1371-1375 (1989)

- Bell S. C., Secretory endometrial and decidual proteins: studies and clinical significance of a maternally derived group of pregnancy-associated serum proteins. Human Reproduction 1, 129-143 (1986)

- Bennett A. L., Paulson K. E., Miller R. E., and Darnell J. E., Aquisition of antigens characteristic of adult pericentral hepatocytes by differentiating fetal hepatoblasts *in vitro*. J. Cell Biol. 105, 1073-1085 (1987)

- Bennett K. L., Lalley P. A., Barth R. K., and Hastie N. D., Mapping the structural genes coding for the major urinary proteins in the mouse: combined use of recombinant inbred strains and somatic cell hybrids. Proc. Natl. Acad. Sci. USA. 79, 928-37 (1982)

- Berendsen H. J. C., Postma J. P. M., van Gunsteren W. F., DiNola A., and Haak J. R., Molecular dynamics with coupling to an external heat bath. J. Chem. Phys. 81, 3684-3690 (1984)

- Berman P., Gray P., Chen E., Keyser K., Erhlich D., Karten H., LaCorbiere M., Esch F., and Schubert D., Sequence analysis, cellular localization, and expression of a neuroretina adhesion and cell survival molecule. Cell 51, 135-142 (1987)

- Bernard A. M., Lauwerys R. R., Noël A., Vandeleene B., and Lambert A., Urine Protein 1: a sex-dependent marker of tubular or glomerular dysfunction. Clinical Chemistry 35, 2141-2442 (1989)

- Bernstein F. C., Koetzle T. F., Williams G. J. B., Meyer (Jr.) E. F., Brice M. D., Rodgers J. R., Kennard O., Shimanouchi T., and Tasumi M., The Protein Data Bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535-542 (1977)

- Bignetti E., Tirindelli R., and Rossi G. L., Crystallization of an odorant-binding protein from cow nasal mucosa. J. Mol. Biol. 186, 211-212 (1985)

- Bishop J. O., Clark A. J., Clissold P. M., Mainey S., and Francke U., Two main groups of mouse major urinary protein genes, both largely located on Chromosome 4. EMBO J. 1, 615-620 (1982)

- Blundell T. L., and Johnson L. N., Protein crystallography. Academic Press, New York (1976)

- Blundell T. L., Sibanda B. L., Sternberg M. J. E., and Thornton J. M., Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326, 347-352 (1987)

- Borghoff S. J., Strasser J., Charbonneau M., and Swenberg J. A., Analysis of 2,4,4-trimethyl-2-pentanol (TMP-OH) binding to male rat kidney α_{2u}-globulin (α_{2u}) and other proteins. Toxicologist 8, 135 (1988)

- Borghoff S. J., Miller A. B., Bowen J. P., and Swenberg J. A., Characteristics of Chemicals Binding to α_{2u}-Globulin *in vitro* - Evaluating Structure-Activity Relationships. Toxicol. Appl. Pharmacol. 107, 228-238 (1991)

- Böcskei Zs., Findlay J. B. C., North A. C. T., Phillips S. E. V., Somers W. S., Wright C. E., Lionetti C., Tirindelli R., and Cavaggioni A., Crystallization of and preliminary X-ray data for the mouse major urinary protein and rat α-2u globulin. J. Mol. Biol. 218, 699-701 (1991)

- Bourne P. E., and Hendrickson W. A., A CPU benchmark for protein crystallographic refinement. Comput. Biol. Med. 20, 219-230 (1990)

- Bowie J. U., Lüthy R., and Eisenberg D., A method to identify protein sequences that fold into a known three-dimensional structure. Science 253, 164-170 (1991)

- Bradford M. M., Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. Analyt. Biochem. 72, 248-254 (1976)

- Braunitzer G., Chen R., Schrank B., and Stangl A., The automatical sequence analysis of a protein (β-lactoglobulin AB). Hoppe-Seyler's Z. Physiol. Chem. 353, 832-834 (1972)

- Brooks B. R., Bruccoleri R. E., Olafson B. D., States D. J., Swaminathan S., and Karplus M., CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 4, 187-217 (1983)

- Brooks D. E., Means A. R., Wright E. J., Sing S. P., and Tiver K. K., Molecular cloning of the cDNA for two major androgen-dependent secretory proteins of 18.5 kilodaltons synthesized by the rat epididymis. J. Biol. Chem. 261, 4956-4961 (1986)

- Brugè F., and Fornili S. L., On the systolic calculation of all-pairs interactions using transputer arrays. J. Comp. Physics 96, 244-228 (1991).

- Brugè F., Martorana V., and Fornili S. L., Molecular Simulation 1, 309 (1988).

- Brünger A. T., X-PLOR: A system for crystallography and NMR. Yale University, New Haven (1990)

- Brünger A. T., Kuriyan J., and Karplus M., Crystallographic R factor refinement by molecular dynamics. Science 235, 458-460 (1987)

- Cancedda F. D., Manduca P., Tacchetti C., Fossa P., Quarto R., and Cancedda R., Developmentally regulated synthesis of a low molecular weight protein (Ch 21) by differentiating chondrocytes. J. Cell Biol. 107, 2455-2463 (1988)

- Cancedda F. D., Asaro D., Molina F., Cancedda R., Caruso C., Camardella L., Negri A., and Ronchi S., The amino terminal sequence of the developmentally regulated Ch21 protein shows homology with amino terminal sequences of low molecular weight proteins binding hydrophobic molecules. Biochem. Biophys. Res. Comm. 168, 933-938 (1990)

- Cavaggioni A., Sorbi R. T., Keen J. N., Pappin D. J. C., and Findlay J. B. C., Homology between the pyrazine-binding protein from nasal mucosa and major urinary proteins. FEBS Lett. 212, 225-228 (1987)

- Cavaggioni A., Findlay J. B. C., and Tirindelli R., ligand-binding characteristics of homologous rat and mouse urinary proteins and pyrazine-binding protein of calf. Comp. Biochem. Physiol. 96B, 513-520 (1990)

- Chamberlain R. M., Gray codes, Fast Fourier Transforms and hypercubes. Parallel Computing 6, 225-233 (1988)

- Chan Y. L., Paz V., and Wool I. G., The primary structure of rat α_{2u} globulin-related protein. Nucleic Acids Research 16, 11368 (1988)

- Charbonneau M., Lock E. A., Strasser J., Cox M. G., Turner M. J., and Bus J. S., 2,2,4-Trimethylpentane induced Nephrotoxicity. I. Metabolic disposition of TMP in male and female Fischer 344 rats. Toxicol. Appl. Pharmacol. 91, 171-181 (1987)

- Charbonneau M., and Swenberg J. A., Studies on the biochemical mechanism of α_{2u}-globulin nephropathy in rats. Chemical Industry Institute of Toxicology Activities 8, 1-5 (1988)

- Charbonneau M., Strasser J., Lock E. A., Turner M. J., and Swenberg J. A., 1,4-Dichlorobenzene (1,4-DCB) - induced nephrotoxicity: Similarity with unleaded gasoline (UG) - induced renal effects. In Nephrotoxicity: Extrapolation from *in vitro* to *in vivo* and from Animal to Man. Bach P. H., and Lock E. A., (eds.) Plenum, New York, (1988a)

- Charbonneau M., Strasser J., Borghoff S. J., and Swenberg J. A., *In vitro* hydrolysis of [^{14}C]-α_{2u}-globulin (α_{2u}) isolated from male rat kidney. Toxicologist 8, 135 (1988b)

- Chatterjee B., Demyan W. F., Song C. S., Garg B. D., and Roy A. K., Loss of androgenic induction of α_{2u}-globulin gene family in the liver of NIH black rats. Endocrinology 125, 1385-1388 (1989)

- Cho Y., Batt C. A., and Sawyer L., Probing the retinol binding site of bovine β-lactoglobulin. To be submitted (1992)

- Chothia C., and Lesk A. M., The relation between the divergence of sequence and structure in proteins. EMBO J. 5, 823-826 (1986)

- Chothia C., Principals that determine the structure of proteins. Ann. Rev. Biochem. 53, 537-572 (1984)

- Claessens M., van Cutsem E., Lasters I., and Wodak S., Modelling the polypeptide backbone with 'spare parts' from known protein structures. Protein Engineering 2, 335-345 (1989)

- Clark A. J., Clissold P. M., Al-Shawi R., Beattie P., and Bishop J., Structure of mouse major urinary protein genes: different splicing configurations in the 3'-non-coding region. EMBO J. 3, 1045-1052 (1984)

- Clore G. M., Driscoll P., Wingfield P. T., and Gronenborg A. M., Low resolution structure of interleukin-1β in solution derived from ^{15}N-^{1}H heteronuclear three-dimensional nuclear magnetic resonance spectroscopy. J. Mol. Biol. 214, 811-817 (1990)

- Colantuoni V., Romano V., Bensi G., Santoro C., Costanzo F., Raugei G., and Cortese R., Cloning abd sequencing of a full length cDNA coding for human retinol binding protein. Nucleic Acids Res. 11, 7769-7776 (1983)

- Connolly M. L., Computation of molecular volume. J. Am. Chem. Soc. 107, 1118-1124 (1985)

- Cooley J. W., and Tukey J. W., An algorithm for the machine calculation of complex Fourier series. Math. Comput. 19, 297-301 (1965)

- Cowan S. W., Newcomer M. E., and Jones T. A., Crystallographic refinement of human serum retinol binding protein at 2Å resolution. PROTEINS: Structure, Function, and Genetics 8, 44-61 (1990)

- Csanalosi I., Schweizer E., Case W. G., and Rickels K., Gepirone in anxiety - a pilot study. J. Clin. Psychopharmacol. 7, 31-33 (1987)

- Dayhoff M., Barker W. C., and Hunt L. T., Establishing homologies in protein sequences. Methods Enzymol. 91, 524-545 (1983)

- Desmet J., Maeyer M. D., Hazes B., and Lasters I., The dead-end elimination theorem and its use in protein side-chain positioning. Nature 356, 539-542 (1992)

- Dettmer R., The artful transputer. Electronics and Power, pg 578 (1986)

- Devereux J., Haeberli P., and Smithies O., A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Research 12, 387-395 (1984)

- Diamond R., Real Space Refinement. In Methods in Enzymology 115, Wyckoff H. W., Hirs C. H. W., and Timasheff S. N. (eds.) pg 237-252 (1985)

- Dice J. F., Molecular determinants of protein half-lives in eukaryotic cells. FASEB J. 1, 349-357 (1987)

- Dietrich D. R., and Swenberg J. A., The presence of α_{2u}-globulin is necessary for *d*-limonene promotion of male rat kidney tumors. Cancer Research 51, 3512-3521 (1991a)

- Dietrich D. R., and Swenberg J. A., NCI-Black-Reiter (NBR) male rats fail to develop renal disease following exposure to agents that induce α_{2u}-globulin (α_{2u}-G) nephropathy. Fundam. Appl. Toxicol. 16, 749-762 (1991b)

- Dolan K. P., Unterman R., McLaughlin M., Nakhasi H. L., Lynch K. R., and Feigelson, P., The structure and expression of a very closely related member of the α_{2u} globulin gene family. J. Biol. Chem. 257, 13527-13534 (1982)

- Dourish C. T., Hutson P. H., Kennett G. A., and Curzon G., 8-OH-DPAT-induced hyperphagia: its neural basis and possible therapeutic relevance. Appetite 7 (Suppl.) 127-140 (1986)

- Drayna D., Fielding C., McLean J., Baer B., Castro G., Chen E., Comstock L., Henzel W., Kohr W., Rhee L., Wion K., and Lawn R., Cloning and expression of human apolipoprotein D cDNA. J. Biol. Chem. 261, 16535-16539 (1986)

- Drickamer K., Kwoh T. J., and Kurtz D. T., Amino acid sequence of the precursor of rat liver α_{2u}-globulin. J. Biol. Chem. 256, 3634-3636 (1981)

- Ducriux A., and Giegé R., (eds.) Crystallization of nucleic acids and proteins *A practical approach.* IRL Press, Oxford (1992)

- Dufour E., and Haertle T., Binding affinities of beta-ionone and related flavor compounds to beta-lactoglobulin - effects of chemical modifications. J. Agric. Food Chem. 38, 1691-1695 (1990)

- Dufour E., Marden M. C., and Haertle T., β-Lactoglobulin binds retinol and protoporphyrin IX at two different binding sites. FEBS Lett. 277, 223-226 (1991)

- Eisenberg D., and McLachlan A. D., Solvation energy in protein folding and binding. Nature 319, 199-203 (1986)

- Ekstrom B., Peterson P. A., and Berggard I., A urinary and plasma α_1-glycoprotein of low molecular weight: Isolation and some properties. Biochem. Biophys. Res. Comm. 65, 1427-1433 (1975)

- Escribano J., Grubb A., and Mendez E., Identification of retinol as one of the protein HC chromophores. Biochem. Biophys. Res. Comm. 155, 1424-1429 (1988)

- Escribano J., Lopex-Otin C., Hjerpe A., Grubb A., and Mendez E., Location and characterization of the three carbohydrate prosthetic groups of human protein HC. FEBS Lett 266, 167-170 (1991)

- Fex G., Albertsson P.-Å., and Hansson B., Interaction between prealbumin and retinol binding protein studied by affinity chromatography, gel filtration and two-phase partition. Eur. J. Biochem. 99, 353-360 (1979)

- Fielding P. E., and Fielding C. J., A cholesteryl ester transfer complex in human plasma. Proc. Natl. Acad. Sci. USA 77, 3327-3330 (1980)

- Fincham D., and Mitchell P. J., Multicomputer molecular dynamics simulation using distributed neighbour lists. Daresbury preprint, Daresbury Laboratory, UK. (1990)

- Finlayson J. S., Mushinski J. F., Hudson D. M., and Potter M., Components of the Major Urinary Protein Complex of Inbred Mice: Separation and Peptide Mapping. Biochem. Genet. 2, 127-140 (1968)

- Finlayson J. S., Potter M., Shinnick C. S., and Smithies O., Components of the Major Urinary Protein Complex of Inbred Mice: Determination of NH_2-Terminal Sequences and Comparison with Homologous Components from Wild Mice. Biochem. Genet. 11, 325-335 (1974)

- Fischer G., Wittman-Liebold B., Lang K., Kiefhaber T., and Schmid F. X., Cyclophilin and peptidyl-prolyl *cis-trans* isomerase are probably identical proteins. Nature 337, 476-478 (1989)

- Flannery A. V., Dalzell G. N., and Beynon R. J., Proteolytic activity in mouse urine - relationship to the kidney metalloendopeptidase, meprin. Biochimica et Biophysica Acta 1041, 64-70 (1990)

- Flower D. R., PhD. Thesis, Astbury Department of Biophysics, The Univesity of Leeds. (1992)

- Fourman J., and Moffat D. B., The blood vessels of the kidney. Blackwell Scientific, Oxford and Edinburgh (1971)

- Fuentealba I. C., Haywood S., and Foster J., Cellular mechanisms of toxicity and tolerance in the copper-loaded rat. III Ultrastructural changes and copper localization in the kidney. Br. J. Exp. Path. 70, 543-556 (1989)

- Fujinaga M, and Read R. J., Experiences with a new translation-function program. Acta Cryst. 20, 517-521 (1987)

- Garavito R., M., Rossmann M. G., Argos P., and Eventoff W., Convergence of active center geometries. Biochemistry 16, 5065-5071 (1977)

- Gaworski C. L., MacEwen J. D., Vernot E. H., Bruner R. H., and Cowan M. J., Comparison of the subchronic inhalation toxicity of petroleum and oil shale JP-5 jet fuels. In The Toxicology of Petroleum Hydrocarbons, MacFarland H. N., Holdworth C. E., MacGregor J. A., Call R. W., and Kane M. L., eds., pg 67-75. American Petroleum Institute, Washington, D.C. (1982)

- Ghosh D., Weeks C. M., Grochulski P., Duax W. L., Erman M., Rimsay R. L., and Orr J. C., Three-dimensional structure of holo 3α,20β-hydroxysteriod dehydrogenase: A member of a short-chain dehydrogenase family. Proc. Natl. Acad. Sci. USA 88, 10064-10068 (1992)

- Gibson J. E., and Bus J. S., Current perspectives in gasoline (light hydrocarbon) induced male rat nephropathy. Ann. N. Y. Acad. Sci. 534, 481-485 (1988)

- Giguere V., Ong E. S., Segui P., and Evans R. M., Identification of a receptor for the morphogen retinoic acid. Nature 330, 624-629 (1987)

- Gilmore C. J., Henderson A. N., and Bricogne G., A multisolution method of phase determination by combined maximization of entropy and likelihood. V. The use of likelihood as a discriminator of phase sets produced by the SAYTAN program for a small protein. Acta Cryst. A47, 842-845 (1991)

- Glennon R. A., Naiman N. A., Lyon R. A., and Titeler M., Arylpiperazine derivatives as high-affinity 5-HT_{1A} serotonin ligands. J. Med. Chem. 31, 1968-1971 (1988a)

- Glennon R. A., Naiman N. A., Pierson M. E., Titeler M., Lyon R. A., and Weisberg E., NAN-190: an arlypiperazine analog that antagonizes the stimulus effects of the 5-HT_{1A} agonist 8-hydroxy-2-(di-*n*-propylamino) tetralin (8-OH-DPAT). European Journal of Pharmacology 154, 339-341 (1988b)

- Godovac-Zimmermann J., The structural motif of β-lactoglobulin and retinol binding protein: a basic framework for binding and transport of small hydrophobic molecules? Trends in Biochem. Sci. 13, 64-66 (1988)

- Gould R. O., and Taylor P., CALC: Interactive program for molecular geometry. University of Edinburgh, Scotland (1983)

- Green D. W., and Aschaffenberg R., Twofold symmetry of the β-lactoglobulin molecule in crystals. J. Mol. Biol. 1, 54-64 (1959)

- Greer J., Comparative modeling methods: Applications to the family of the mammalian serine proteases. PROTEINS: Structure, Function, and Genetics 7, 317-334 (1990)

- Greengard L., and Gropp W. D., A parallel version of the fast multipole method. Comp. Math. Appl. 20, 63-71 (1990)

- Gros P., van Gunsteren W. F., and Hol W. G. J., Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. Science 249, 1149-1152 (1990)

- Grubmüller H., Heller H., Windemuth A., and Schulten K., Generalized Verlet algorithm for efficient molecular dynamics simulations with long-range interactions. Molecular Simulation 6 (1991)

- Haars L. J., and Pitot H. C., Hormonal and developmental regulation of glycosylated α_{2u}-globulin synthesis. Archives of Biochemistry and Biophysics. 201, 556-563 (1980)

- Halder C. A., Warne T. M., and Hatoum N. S., Renal toxicity of gasoline and related petroleum napthas in male rats. In Proceedings of the Workshop on the Kidney Effects of Hydrocarbons, Hemstreet G. P., Thorpe J. J., and Kane M. L., eds. Pg 107-125. American Petroleum Institute, Washington, D.C. (1983)

- Halder C. A., Holdsworth C. E., Cockrell B. Y., and Piccirillo V. J., Hydrocarbon nephropathy in male rats. Identification of the nephropathic components of unleaded gasoline. Toxicol. Ind. Health 1, 67-87 (1985)

- Halliday J. A., Bell K., and Shaw D. C., The complete amino acid sequence of feline β-lactoglobulin II and a partial revision of the equine β-lactoglobulin II sequence. Biochimica et Biophysica Acta 1077, 25-30 (1991)

- Hambling S. G., McAlpine A. S., and Sawyer L., β-lactoglobulin. In Milk Proteins, Fox P. (ed.), in press, Elsevier publications (1991)

- Handshumacher R. E., Harding M. W., Rice J., Drugge R. J., and Speicher D. W., Cyclophilin: A specific cytosolic binding protein for cyclosporin A. Science 226, 544-547 (1984)

- Hansch G. M., The homologous species restriction of the complement attack: Structure and function of the C8 binding protein. Current Topics in Microbiology and Immunology 140, 109-118 (1988)

- Härd T., Kellenback E., Boelens R., Maler B. A., Dahlman K., Freedman L. P., Carlstedt-Duke J., Yamamoto K. R., Gustafsson J.-A., and Kaptein R., Solution structure of the glucocorticoid receptor DNA-binding domain. Science 249, 157-160 (1990)

- Hastie H. D., Held W. A., and Toole J. J., Multiple genes coding for the androgen-regulated major urinary proteins of the mouse. Cell 17, 449-457 (1979)

- Havel T. F., and Snow M. E., A new method for building protein conformations from sequence alignments with homologues of known structure. J. Mol. Biol. 217, 1-7 (1991).

- Hayaishi O., Sleep-wake regulation by prostaglandins D_2 and E_2. J. Biol. Chem. 263, 14593-14596 (1988)

- Heller H., Grubmüller H., and Schulten K., Molecular dynamics simulation on a parallel computer. Molecular Simulation 5, 133-165 (1990)

- Hendlich M., Lackner P., Weitckus S., Floeckner H., Froschauer R., Gottsbacher K., Casari G., and Sippl M. J., Identification of native protein folds amongst a large number of incorrect models: The calculation of low energy conformations from potentials of mean force. J. Mol. Biol. 216, 167-180 (1990)

- Hendrickson W. A., Stereochemically restrained refinement of macromolecular structures. In Methods in Enzymology 115, Wyckoff H. W., Hirs C. H. W., and Timasheff S. N. (eds.) pg 252-270 (1985)

- Hendrickson W. A., Pahler A., Smith J. L., Satow Y., Merritt E. A., and Phizackerley R. P., Crystal structure of core streptavidin determined from

multiwavelength anomalous diffraction of synchrotron radiation. Proc. Natl. Acad. Sci. USA 86, 2190-2194 (1989)

- Henzel W. J., Rodriguez H., Singer A. G., Stults J. T., Macrides F., Agosta W. C., and Niall H., The primary structure of aphrodisin. J. Biol. Chem. 263, 16682-16687 (1988)

- Higgins D. G., and Sharp P. M., Fast and sensitive multiple sequence alignments on a microcomputer. CABIOS 5, 151-153 (1989)

- Hirer B. C., Roth H. L., and Peroutka S. J., Antimigraine drug interactions with 5-hydroxytryptamine$_{1A}$ receptors. Ann. Neurol. 19, 511-513 (1986)

- Hoare C. A. R., Communicating Sequential Processes. Prentice Hall International, Hemel Hempstead (1985)

- Hockney R. W., and Jessop C. R., Parallel Computers. Adam Hilger Ltd., Bristol (1983)

- Holden H. M., Rypniewski W. R., Law J. H., and Rayment I., The molecular structure of insecticyanin from the tobacco hornworm *Manduca sexta* L. at 2.6 Å resolution. EMBO J. 6, 1565-1570 (1987)

- Hraba-Renevey S., Turler H., Kress H., Salomom C., and Weil R., SV40-induced expression of mouse gene 24p3 involves a post-transcriptional mechanism. Oncogene 4, 601-608 (1989)

- Hubbard R. E., and Baker E. N., Hydrogen bonding in globular proteins. Prog. Biophys. Molec. Biol. 44, 97-179 (1984)

- Hubbard R. E., In Computer aided molecular design. pg 99-106, Oyez Scientific, UK. (1985)

- Huber R., Experience with the application of Patterson search techniques. In Molecular Replacement, Proceedings of the Daresbury Study Weekend, Machin P. A. (eds.) pg 58-61 (1985)

- Huber R., Schneider M., Epp O., Mayr I., Messerschmidt A., and Pflugrath J., Crystallization, crystal structure analysis and preliminary molecular model of the bilin binding protein from the insect *Pieris brassicae.* J. Mol. Biol. 195, 423-434 (1987a)

- Huber R., Schneider M., Mayr I., Muller R., Deutzmann R., Suter F., Zuber H., Falk H., and Kayser H., Molecular structure of the bilin binding protein (BBP) from *Pieris brassicae* after refinement at 2.0 Å resolution. J. Mol. Biol. 198, 499-513 (1987b)

- INMOS Limited Transputer Reference Manual. Prentice Hall International, Hemel Hempstead (1988)

- International Tables for X-ray Crystallography, Volume IV, pg 99-102, Kynoch Press, Birmingham (1974)

- Jack A., and Levitt M., Refinement of large structures by simultaneous minimization of energy and R factor. Acta Cryst. A34, 931-935 (1978)

- Jacobsen N. O., and Jørgensen F., Ultrastructural observations on the *pars decendens* of the proximal tubule in the kidney of the male rat. Z. Zellforch 136, 479-499 (1973)

- Jakoby W. B., Bend J. R., and Caldwell J., eds. Metabolic Basis of Detoxication: Metabolism of Functional Groups. pg 8-14 and 91-97. Academic press, N. Y. (1982)

- Janin J., and Chothia C., FEBS: 12th Meeting Dresden 1978. Hofmann E., Pfeil W., and Aurich H. (eds.) 52, 227-237, Pergammon, Oxford (1979)

- Janin J., Surface area of globular proteins. J. Mol. Biol. 105, 13-14 (1976)

- Janin J., Surface and inside volumes in globular proteins. Nature 277, 491-492 (1979)

- Jones D. T., Taylor W. R., and Thornton J. M., A new approach to protein fold recognition. Nature 358, 86-89 (1992)

- Jones G., and Goldsmith M., Programming in Occam 2. Prentice Hall International, Hemel Hempstead (1988)

- Jones T. A., A graphics model building and refinement system for macromolecules. J. Appl. Cryst. 11, 268-272 (1978)

- Jones T. A., Bergfors T., Sedzik J., and Unge T., The three-dimensional structure of P2 myelin protein. EMBO J. 7, 1597-1604 (1988)

- Jones T. A., and Kjeldgaard M., Manual for O version 5.7, Department of Molecular Biology, Uppsala, Sweden (1991)

- Julkunen M., Seppala M., and Janne O. A., Complete amino acid sequence of human placental protein 14: A progesterone-regulated uterine protein homologous to β-lactoglobulins. Proc. Natl. Acad. Sci. USA 85, 8845-8849 (1988)

- Kabsch W., and Sander C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22, 2577-2637 (1983)

- Kallen J., Spitzfaden C., Zurini M. G. M., Wider G., Widmer H., Wüthrich K., and Walkinshaw M. D., Structure of human cyclophilin and its binding site for cyclosporin A determined by X-ray crystallography and NMR spectroscopy. Nature 353, 276-279 (1991)

- Kanai M., Raz A., and Goodman D. S., Retinol binding protein: The transport protein for vitamin A in human plasma. J. Clin. Invest. 47, 2025-2044 (1968)

- Karplus M., and Petsko G. A., Molecular dynamics simulations in biology. Nature 347, 631-639 (1990)

- Kimura H., Odani S., Suzuki J., Arakawa M., and Ono T., Kidney fatty acid-binding protein: identification as α_{2u}-globulin. FEBS Lett. 246, 101-104 (1989)

- Kimura H., Odani S., Nishi S., Sato H., Arakawa M., and Ono T., Primary structure and cellular distribution of two fatty acid-binding proteins in adult rat kidneys. J. Biol. Chem. 266, 5963-5972 (1991)

- Kirkpatrick S., Gelatt (Jr.) C. D., and Vecchi M. P., Optimization by simulated annealing. Science 220, 671-680 (1983)

- Kloss M. W., Cox M. G., Norton R. M., Swenberg J. A., and Bus J. S., Sex-dependent differences in the disposition of ^{14}C-5-2,2,4-trimethylpentane in Fischer 344 rats. In Renal Heterogeneity and Target Cell Toxicity. Bach P. H., and Lock E. A., (eds.) pg 489-492. Wiley, Chichester (1985)

- Kollman, P., Shashidhar R., Brown F., Daggett V., Seibel G., and Chandra Singh U., Free energy perturbation methods can give exciting insights into the effect of site-specific mutants on both binding and catalysis: Applications to subtilisin, trypsin and triose phosphate isomerase and the description of a free energy component analysis. In Protein Structure, Folding, and Design 2, pg 215-225, Alan R. Liss Inc. (1987)

- Konnert J. H., and Hendrickson W. A., A restrained-parameter thermal-factor refinement procedure. Acta Cryst. A36, 344-350 (1980)

- Kremer J. M. H., Wilting J., and Janssen L. H. M., Drug binding to human alpha-1-acid glycoprotein in health and disease. Pharmacol. Reviews 40, 1-47 (1988)

- Kuhn N. J., Woodworth-Gutai M., Gross K. W., and Held W. A., Subfamilies of the mouse major urinary protein (MUP) multi-gene family: sequence analysis of cDNA clones and differential regulation in the liver. Nucleic Acids Research 12, 6073-6090 (1984)

- Laemmli U. K., Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 299, 592-596 (1970)

- Lane S. E., and Neuhaus O. W., Further studies on the isolation and characterization of a sex-dependent protein from the urine of male rats. Biochim. Biophys. Acta. 257, 461-470 (1972)

- Lee K. H., Wells R. G., and Reed R. R., Isolation of an olfactory cDNA: Similarity to retinol-binding protein suggests a role in olfaction. Science 235, 1053-1056 (1987)

- Lehman-McKeeman L. D., Rivera-Torres M. I., and Caudill D., Lysosomal degradation of α_{2u}-globulin and α_{2u}-globulin-xenobiotic conjugates. Toxicol. Appl. Pharmacol. 103, 539-548 (1990)

- Lehman-McKeeman L. D., Rodriguez P. A., Caudill D., Fey M. L., Eddy C. L., and Asquith T. N., Hyaline droplet nephropathy resulting from exposure to 3,5,5-trimethylhexanosyloxybenzene sulphonate. Toxicol. Appl. Pharmacol. 107, 429-438 (1991)

- Lehman-McKeeman L. D., and Caudill D., Biochemical basis for mouse resistance to hyaline droplet nephropathy: lack of relevance of the α_{2u}-globulin protein superfamily in this male rat-specific syndrome. Toxicol. Appl. Pharmacol. 112, 214-221 (1992)

- Lesk A. M., Branden, C. I., and Chothia C., Structural principles of α/β barrel proteins: The packing of the interior of the sheet. PROTEINS: Structure, Function, and Genetics 5, 139-148 (1989)

- Lesk A. M., Protein architecture: *A practical approach.* IRL Press, Oxford (1991)

- Lipson H., and Cochran W., The determination of crystal structures. Bell, London (1957)

- Lock E. A., Charbonneau M., Strasser J., Swenberg J. A., and Bus J. S., 2,2,4-Trimethylpentane-induced nephrotoxicity. II. The reversible binding of a TMP metabolite to a renal protein fraction containing α_{2u}-globulin. Toxicol. Appl. Pharmacol. 91, 182-192 (1987)

- Logothetopoulus J., and Weinbren K., Naturally occurring protein droplets in the proximal tubule of the rat's kidneys. Brit. J. Exp. Path. 36, 402-409 (1955)

- Lorusso J. R., Moffat S., and Ohman J. L., Immunologic and biochemical properties of the major mouse urinary allergen (Mus m I). J. Allergy Clin. Immunol. 78, 928-37 (1986)

- Loury D., Smith-Oliver T., and Butterworth B. E., Assessment of the binding potential of 2,2,4-trimethylpentane to the rat α_{2u}-globulin. Toxicol. Appl. Pharmacol. 88, 44-56 (1987)

- Lüthy R., Bowie J. U., and Eisenberg D., Assessment of protein models with three-dimensional profiles. Nature 356, 83-85 (1992)

- MacFarland N. H., Xenobiotic Induced Kidney Lesions: Hydrocarbons. The PS-6 90-day and 2-year gasoline studies. In Proceedings of the workshop on the kidney effects of hydrocarbons. Hemstreet G. P., Thorpe J. J., and Kane M. L., (eds.) pg 68-74. American Petroleum Institute, Washington, D.C. (1983)

- MacNaughton M. G., and Uddin D. E., Toxicology of mixed distillate and high energy synthetic fuels. In Proceedings of the workshop on the kidney effects of hydrocarbons. Hemstreet G. P., Thorpe J. J., and Kane M. L., (eds.) pg 171-185. American Petroleum Institute, Washington, D.C. (1983)

- Mandlebrot B. B., The Fractal Geometry of Nature. Freeman (1982)

- Margolis F. L., Olfactory receptor neurons: specific gene expression and a hypothetical model for stimulus receptors. Discussions in Neurosciences 4, 47-52 (1987)

- Martin G. E., and Lis (Jr) E. V., Hypotensive action of 8-hydroxy-2-(di-N-propylamino) tetralin (8-OH-DPAT) in spontaneously hypertensive rats. Arch. Int. Pharmacodyn. 273, 251-261 (1985)

- Matuo Y., Nishi N., Tanaka Y., Muguruma Y., Tanaka K., Akatsuka Y., Matsui S. I., Sanberg A., and Wada F., Changes of an androgen-dependent nuclear protein during functional differentiation and by dedifferentiation of the dorsolateral prostate of rats. Biochem. Biophys. Res. Comm. 118, 467-473 (1984)

- McCammon J. A., and Harvey S. C., Dynamics of proteins and nucleic acids. Cambridge University Press, Cambridge (1987)

- McConathy W. J., and Alaupovic P., Isolation and partial characterization of apolipoprotein D: A new protein moiety of the human plasma lipoprotein system. FEBS Lett. 37, 178-182 (1973)

- McPherson A., Preparation and analysis of protein crystals. Wiley, New York (1982)

- McRee D. E., Meyer T. E., Cusanovich M. A., Parge E., and Getzoff E. D., Crystallographic characterization of a photoactive yellow protein with photochemistry similar to sensory rhodopsin. J. Biol. Chem. 261, 13850-13851 (1986)

- McRee D. E., Tainer J. A., Meyer T. E., van Beeumen J., Cusanovich M. A., and Getzoff E. D., Crystallographic structure of a photoreceptor protein at 2.4 Å resolution. Proc. Natl. Acad. Sci. USA 86, 6533-6537 (1989)

- Messerschmidt A., and Pflugrath J. W., Crystal orientation and X-ray pattern prediction routines for area-detector diffractometer systems in macromolecular crystallography. J. Appl. Cryst. 20, 306-315 (1987)

- Metcalf M., and Reid J., Fortran 90 Explained. Oxford Scientific Publications (1990)

- Michmick S. W., Rosen M. K., Wandless T. J., Karplus M., and Schreiber S. L., Solution structure of FKBP, a rotamase enzyme and receptor for FK506 and rapamycin. Science 252, 836-839 (1991)

- Monaco H. L., Zanotti G., Spadon P., Bolognesi M., Sawyer L., and Eliopoulos E. E., Crystal structure of the trigonal form of bovine beta-lactoglobulin and of its complex with retinol at 2.5 Å resolution. J. Mol. Biol. 197, 695-706 (1987)

- Muller-Eberhard J., The membrane attack complex of complement. Ann. Rev. Immunol. 4, 503-528 (1986)

- Muller-Fahrnow A., Egner U., Jones T. A., Rudel H., Spener F., and Saenger W., Three-dimensional structure of fatty-acid-binding protein from bovine heart. Eur. J. Biochem. 199, 271-276 (1991)

- Murthy M. R., Reid T. J., Sicignano A., Tanaka N., and Rossmann M. G., Structure of beef liver catalase. J. Mol. Biol. 152, 465-499 (1981)

- Muskett F. W., A study of phosphoprotein and phospopeptide interactions with calcium ions and dicalcium phosphate dihydrate crystals. PhD Thesis, Department of Biochemistry, University of Edinburgh (1992)

- Nagata A., Suzuki Y., Igarashi M., Eguchi N., Toh H., Urade Y., and Hayaishi O., Human brain prostaglandin D synthase has been evolutionarily differentiated from lipophilic-ligand carrier proteins. Proc. Natl. Acad. Sci. USA. 88, 4020-4024 (1991)

- Needleman S. B., and Wunsch C. D., A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453 (1970)

- Neuhaus O. W., Flory W., Biswas N., and Hollerman C. E., Urinary excretion of α_{2u}-globulin and albumin by adult male rats following treatment with nephropathic agents. Nephron 28, 133-140 (1981)

- Newcomer M. E., and Ong D. E., Purification and crystallization of a retinoic acid-binding protein from rat epididymis. J. Biol. Chem. 265, 12876-12879 (1990)

- Newcomer M. A., Liljas A., Sundelin J., Rask L., and Peterson P. A., Crystallization of and preliminary X-ray data for the plasma retinol-binding protein. J. Biol. Chem. 259, 5230-5231 (1984)

- Nilges M., and Brünger A. T., Automated modeling of coiled coils: application to the GCN4 dimerization region. Protein Engineering 4, 649-659 (1991)

- North A. C. T., Phillips D. C., and Mathews F. S., A semi-empirical method of absorption correction. Acta Cryst. A24, 351 (1968)

- Novotny J., Bruccoleri R., and Karplus M., An analysis of incorrectly folded protein models: Implications for structural predictions. J. Mol. Biol. 177, 787-818 (1984)

- Novotny J., Rashin A. A., and Bruccoleri R. E., Criteria that discriminate between native proteins and incorrectly folded models. PROTEINS: Structure, Function, and Genetics 4, 19-30 (1988)

- Ockner R. K., Historic overview of studies on fatty acid-binding proteins. Mol. Cell. Biochem. 98, 3-9 (1990)

- Oliver J., and MacDowall M., Cellular mechanisms of protein metabolism in the nephron, I. The structural aspects of proteinuria: tubular absorption, droplet formation, and the disposal of proteins. J. Exp. Med. 99, 589-604 (1954)

- Oliver J., and MacDowall M., Cellular mechanisms of protein metabolism in the nephron, VII. The characteristics and significance of the protein absorption droplets (hyalin droplets) in epidemic hemorrhagic fever and other renal diseases. J. Exp. Med. 107, 731-754 (1958)

- Olson C. T., Yu K. O., Hobson D. W., and Serve M. P., Identification of urinary metabolites of the nephrotoxic hydrocarbon 2,2,4-trimethylpentane in male rats. Biochem. Biophys. Res. Commun. 130, 313-316 (1985)

- Omichinski J. G., Clore G. M., Appella E., Sakaguchi K., and Gronenborn A. M., High resolution three-dimensional structure of a single zinc finger from human enhancer binding protein in solution. Biochemistry 29, 9324-9334 (1990)

- Orbell J. D., Guddat L. W., Machin K. J., and Isaacs N. W., The effect of Fast Protein Liquid Chromatography (FPLC) cationic exchange purification on the crystallization of a monoclonal Fab fragment. Analytical Biochemistry 170, 390-392 (1988)

- Ostlund N. S., and Whiteside R. A., A machine architecture for molecular dynamics: the systolic loop. In Macromolecular structure and specificity: computer-assisted modeling and applications. Venkataraghavan B., and Feldmann R. J. (eds.) pg 195-208. Anals of the New York Academy of Sciences, The New York Academy of Sciences, New York, (1985)

- Papiz M. Z., Sawyer L., Eliopoulos E. E., North A. C. T., Findlay J. B. C., Sivaprasadarao R., Jones T. A., Newcomer M. E., and Kraulis P. J., The structure of β-lactoglobulin and its similarity to plasma retinol-binding protein. Nature 324, 383-385 (1986)

- Paterson G. J., The expression and secretion of ovine β-lactoglobulin in yeast. PhD. Thesis, Department of Biochemistry, University of Edinburgh (1991)

- Peitsch M. C., and Boguski M. S., Is Apolipoprotein D a mammalian bilin binding-protein? New Biologist 2, 1-10 (1990)

- Peitsch M. C., and Boguski M. S., The first lipocalin with enzymatic activity. Trends in Biochem. Sci. 16, 363 (1991)

- Peroutka S. J., Heuring R. E., Mauk M. D., and Kocsis J. D., Analysis of 5-HT1 binding site subtypes and potential functional correlates. Psychopharmacol. Bull. 22, 813-817 (1986)

- Pervais S., and Brew K., Homology and structure-function correlations between α_1-acid glycoprotein and serum retinol-binding protein and its relatives. FASEB J. 1, 209-214 (1987)

- Pevsner J., Trifiletti R. R., Strittmatter S. M., and Snyder S. H., Isolation and characterization of an olfactory receptor protein for odorant pyrazines. Proc. Natl. Acad. Sci. USA 82, 3050-3054 (1985)

- Pevsner J., Reed R., Feinstein P. G., and Snyder S. H., Molecular cloning of odorant-binding protein: member of a ligand carrier family. Science 241, 336-339 (1988)

- Phillips R. D., and Cockrell B. Y., Effect of certain light hydrocarbons on kidney function and structure in male rats. In Proceedings of the workshop on the kidney effects of hydrocarbons. Hemstreet G. P., Thorpe J. J., and Kane M. L., (eds.) pg 130-150. American Petroleum Institute, Washington, D.C. (1983)

- Phillips S. C., A review of the human kidney effects of hydrocarbon exposure. In Proceedings of the workshop on the kidney effects of

hydrocarbons. Hemstreet G. P., Thorpe J. J., and Kane M. L., (eds.) pg 298-318. American Petroleum Institute, Washington, D.C. (1983)

- Ponder J. W., and Richards F. M., Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. J. Mol. Biol. 193, 775-791 (1987)

- Raine A. R. C., Fincham D., and Smith W., Systolic loop methods for molecular dynamics simulation using multiple transputers. Computer Physics Communications 55, 13-20 (1989)

- Raine A. R. C., Systolic loop methods for molecular dynamics simulation, generalised for macromolecules. Molecular Simulation 7, 59 (1991)

- Ramachandran G. N., and Sasisekharan V., Conformation of polypeptides and proteins. Advances in Protein Chem. 23, 283-437 (1968)

- Richardson J. S., The anatomy and taxonomy of protein structure. Advances in Protein Chemistry 34, 167-339 (1981)

- Richmond T. J., Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. J. Mol. Biol. 178, 63-89 (1984)

- Riley C. T., Barbeau B. K., Keim P. S., Kezdy F. J., Heinrikson R. L., and Law J. H., The covalent protein structure of insecticyanin, a blue biliprotein from the hemolymph of the Tobacco Hornworm, *Manduca sexta* L. J. Biol. Chem. 259, 13159-13165 (1984)

- Rogers S., Wells R., and Rechsteiner M., Amino acid sequences common to rapidly degraded proteins: The PEST hypothesis. Science 34, 364-368 (1986)

- Rossman M. G., The molecular replacement method. Gordon and Breach, New York (1972)

- Roy A. K., Neuhaus O. W., and Harrison C. R., Preparation and characterization of a sex-dependent rat urinary protein. Biochimica et Biophysica Acta 127, 72-81 (1966)

- Roy A. K., Androgen-dependent synthesis of the α_{2u}-globulin in the rat: Role of the pituitary gland. J. Endocrind. 56, 295-301 (1973a)

- Roy A. K., Androgenic induction of α_{2u}-globulin in rats: Androgen insensitivity in prepubertal animals. Endocrinology 92, 957-960 (1973b)

- Sacchettini J. C., Gordon J. I., and Banaszak L. J., The structure of crystalline *Escherichia coli*-derived rat intestinal fatty acid-binding protein at 2.5 Å resolution. J. Biol. Chem. 263, 5815-5819 (1988)

- Sander C., and Schneider R., Database of homology-derived protein structures and the structural meaning of sequence alignment. PROTEINS: Structure, Function, and Genetics 9, 56-68 (1991)

- Sanders P. W., and Booker B. B., Pathobiology of cast nephropathy from human Bence-Jones proteins. J. Clin. Invest. 89, 630-639 (1992)

- Sawyer L., and Richardson J. S., Using appropriate nomenclature. Trends in Biochem. Sci. 16, 11 (1991)

- Sawyer L., One fold amongst many. Nature 327, 659 (1987)

- Scapin G., Spadon P., Mammi M., Zanotti G., and Monaco H. L., Crystal structure of chicken liver basic fatty acid-binding protein at 2.7 Å resolution. Mol. Cell. Biochem. 98, 95-99 (1990)

- Schmale H., Holtgreve-Grez H., and Christiansen H., Possible role for salivary gland protein in taste reception indicated by homology to lipophilic-ligand carrier proteins. Nature 343, 366-369 (1990)

- Schmid K., In The Plasma Proteins, Vol. I. Putman F. (ed.) pg 184-228, New York, Academic Press (1975)

- Schreiber H., Steinhauser O., and Schuster P., Parallel molecular dynamics of biomolecules. Parallel Computing 18, 557-573 (1992)

- Schubert D., and LaCorbiere M., Isolation of a cell-surface receptor for chick neural retina adherons. J. Cell Biol. 100, 56-63 (1985)

- Schubert D., LaCorbiere M., and Esch F., A chick neural retina adhesion and survival molecule is a retinol-binding protein. J. Cell Biol. 102, 2295-2301 (1986)

- Scientific Computing Associates, Network Linda. Scientific Computing Associates Inc., New Haven (1992)

- Shanan K., Gilmartin M., and Derman E., Nucleotide Sequences of Liver, Lachrymal, and Submaxillary Gland Mouse Major Urinary Protein mRNA's: Mosiac Structure and Construction of Panels of Gene-Specific Synthetic Oligonucleotide Probes. Mol. Cell. Biol. 7, 1938-1946 (1987a)

- Shanan K., Denaro M., Gilmartin M., Shi Y., and Derman E., Expression of Six Mouse Major Urinary Protein Genes in the Mammary, Parotid, Sublingual, Submaxillary, and Lachrymal Glands and in the Liver. Mol. Cell. Biol. 7, 1947-1954 (1987b)

- Sheldrick G., SHELX86: Program for crystal structure solution. University of Göttingen, Federal Republic of Germany (1986)

- Short B. G., Burnett V. L., and Swenberg J. A., Histopathology and cell proliferation induced by 2,2,4-trimethylpentane in the male rat kidney. Toxicol. Pathol. 14, 194-203 (1986)

- Simard J., Veilleux R., de Launoit Y., Haagensen D. E., and Labrie F., Stimulation of apolipoprotein D secretion by steroids coincides with inhibition of cell proliferation in human LNCaP prostate cancer cells. Cancer Res. 51, 4336-4341 (1991)

- Sixma T. K., Pronk S. E., Kalk K. H., Wartna E. S., van Zanten B. A. M., Witholt B., and Hol W. G. J., Crystal structure of a cholera toxin-related heat-labile enterotoxin from *E. coli.* Nature 351, 371-377 (1991)

- Sjölin L., Prince E., Svensson L. A., and Gilliland G. L., *Ab Initio* Phase determination for X-ray diffraction data from crystals of a native protein. Acta Cryst. A47, 216-223 (1991)

- Snyder S. H., Sklar P. B., and Pevsner J., Molecular mechanisms of olfaction. J. Biol. Chem. 263, 13971-13974 (1988)

- Sodetz J. M., Structure and function of C8 in the membrane attack sequence of complement. Current Topics in Microbiology and Immunology 140, 21-31 (1988)

- Solomon A., Weiss D. T., and Kattine A. A., Nephrotoxic potential of Bence Jones proteins. New England J. Med. 324, 1845-1857 (1991)

- Spence A. M., Sheppard P. C., Davie J. R., Matuo Y., Nishi N., McKeehan W. L., Dodd J. G., and Matusik R. J., Regulation of a bifunctional mRNA results in synthesis of secreted and nuclear probasin. Proc. Natl. Acad. Sci. USA 86, 7843-7847 (1989)

- Sternberg M. J. E., and Islam S. A., Local protein sequence similarity does not imply a structural relationship. Protein Engineering 4, 125-131 (1990)

- Stonard M. D., Phillips P. G. N., Foster J. R., Simpson M. G., and Lock E. A., α_{2u}-Globulin: Measurement in rat kidney following administration of 2,2,4-trimethylpentane. Toxicology 41, 161-168 (1986)

- Stout G. H., and Jensen L. H., X-ray structure determination *A practical guide.* Macmillan Publishing Co. Inc, New York (1968)

- Strasser J., Charbonneau M., Borghoff S. J., Turner M. J., and Swenberg J. A., Renal protein droplet formation in male Fischer 344 rats after isophorone (IPH) treatment. Toxicologist 8, 136 (1988)

- Sugimoto D., Chikada Y., Makino J., Ito T., Ebisuzaki T., and Umemura M., A special-purpose computer for gravitational many-body problems. Nature 345, 33-35 (1990)

- Summers N. L., and Karplus M., Modeling of globular proteins: A distance-based data search procedure for the construction of insertion/deletion regions and Pro $\leftrightarrow$ non-Pro mutations. J. Mol. Biol. 216, 991-1016 (1990)

- Swenberg J. A., Short B., Borghoff S. J., Strasser J., and Charbonneau M. A., The comparative pathobiology of α_{2u}-globulin nephropathy. Toxicol. Appl. Pharmacol. 97, 35-46 (1989)

- Sweny P., Farrington K., and Moorhead J. F., The kidney and its disorders. Blackwell Scientific, Oxford and Edinburgh (1989)

- Szpirer C., Rivière M., Szpirer J., Genet M., Drèze, Islam M. Q., and Levan G., Assignment of 12 loci to rat Chromosome 5: Evidence that this Chromosome is homologous to mouse Chromosome 4 and to human Chromosmes 9 and 1 (1p Arm). Genomics 6, 679-684 (1990)

- Tanford C., Bunville L. G., and Nozaki Y., The reversible transformation of β-lactoglobulin at pH 7.5. J. Am. Chem. Soc. 81, 4032-4036 (1959)

- Taylor G., Experiences in the use of restrained dynamic refinement. In Molecular Simulation and Protein Crystallography, Proceedings of the CCP4 study weekend 27-28 January 1989. Goodfellow J., Hendrick K., and Hubbard R. (eds.), Daresbury Laboratory, UK (1989)

- Tello D., Spinelli S., Souchon H., Saul F. A., Riottot M. M., Mariuzza R. A., Lascombe M. B., Houdusse A., Eisele J. L., Fischmann T., Chitarra V., Boulot G., Bhat T. N., Bentley G. A., Alzari P. M., and Poljak R. J., 3-dimensional structure and antigen-binding specificity of antibodies. Biochimie 72, 507-512 (1990)

- Ten Eyck L. F., Efficient structure-factor calculation for large molecules by the Fast Fourier Transform. Acta Cryst. A33, 486-492 (1977)

- Thinking Machines Corporation, Programming in C/Paris. Thinking Machines Corporation, Cambridge, Massachusetts, USA (1989)

- Thinking Machines Corporation, Connection Machine CM-200 Series Technical Summary. Thinking Machines Corporation, Cambridge, Massachusetts, USA (1991a)

- Thinking Machines Corporation, Programming in CM Fortran. Thinking Machines Corporation, Cambridge, Massachusetts, USA (1991b)

- Thinking Machines Corporation, CM-200 Scientific Software Library Release Notes. Thinking Machines Corporation, Cambridge, Massachusetts, USA (1992)

- Thornton J. M., McArthur M. W., Smith D. K., Gardner S. P., Hutchinson E. G., Morris A. L., and Sibanda B. L., Analysis of errors found in protein structure coordinates in the Brookhaven Data Bank. In Accuracy and reliability of macromolecular crystal structures, Proceedings of the CCP4 study weekend 26-27 January 1990, Henrick K., Moss D. S., and Tickle I. J. (eds.), Daresbury Laboratory, UK (1990)

- Thornton J. M., Flores T. P., Jones D. T., and Swindells M. B., Prediction of progress at last. Nature 354, 105-106 (1991)

- Tirindelli R., Keen J. N., Cavaggioni A., Eliopoulos E. E., and Findlay J. B. C., Complete amino acid sequence of pyrazine-binding protein from cow nasal mucosa. Eur. J. Biochem. 185, 569-572 (1989)

- Tischer C. C., and Brenner B. M., Renal Pathology, Volume 1. Lippincott Company, Philadelphia (1989)

- Tripos Associates, SYBYL: A molecular modelling package. Tripos Associates Inc., St. Louis, Missouri (1991)

- Unterman R. D., Lynch K. R., Nakhasi H. L., Dolan K. P., Hamilton J. W., Cohn D. V., and Feigelson P., Cloning and sequence of several α_{2u}-globulin cDNAs. Proc. Natl. Acad. Sci. USA. 78, 3478-3482 (1981)

- Urade Y., Fujimoto N., and Hayaishi O., Purification and characterization of rat brain prostaglandin D synthetase. J. Biol. Chem. 260, 12410-12415 (1985)

- van der Laan J. M., Swarte M. B. A., Groendijk H., Hol W. G. J., and Drenth J., The influence of purification and protein heterogeneity on the crystallization of *p*-hydroxybenzoate hyroxylase. Eur. J. Biochem. 179, 715-724 (1989)

- Vandoren G., Mertens B., Heyns W., van Baelen H., Rombauts W., and Verhoeven G., Different forms of α_{2u}-globulin in male and female rat urine. Eur. J. Biochem. 134, 175-181 (1983)

- Van Duyne G. D., Standaert R. F., Karplus P. A., Schreiber S. L., and Clardy J., Atomic structure of FKBP-FK506, an immunophilin-immunosuppressant complex. Science 252, 839-842 (1991)

- van Gunsteren W. F., GROMOS: Groningen Molecular Simulation system. BIOMOS b.v., Biomolecular Software, University of Groningen, The Netherlands. (1987)

- van Gunsteren W. F., and Berendsen H. J. C., Algorithms for macromolecular dynamics and constraint dynamics. Molecular Physics 34, 1311-1327 (1977)

- Vingron M., and Argos P., Determination of reliable regions in protein sequence alignments. Protein Engineering 3, 565-569 (1990)

- Wander T. J., Nelson A., Okazaki H., and Richelson E., European J. Pharmacol. 132, 115-121 (1986)

- Weber P. C., Ohlendorf D. H., Wendoloski J. J., and Salemme F. R., Structural origins of high-affinity biotin binding to streptavidin. Science 243, 85-88 (1989)

- Weiner P. K., and Kollman P. A., AMBER: Assisted model building with energy refinement. A general program for modelling molecules and their interactions. J. Comput. Chem. 2, 287-303 (1981)

- Weiner S. J., Kollman P. A., Case D. A., Chandra Singh U., Ghio C., Alagona G., Profeta (Jr.) S., and Weiner P., A new force field for molecular mechanical simulation of nucleic acids and proteins. J. Am. Chem. Soc. 106, 765-784 (1984)

- Weiner S. J., Kollman P. A., Nguyen D. T., and Case D. A., An all atom force field for simulations of proteins and nucleic acids. J. Comput. Chem. 7, 230-252 (1986)

- Windemuth A., and Schulten K., Molecular dynamics simulation on the Connection Machine. Molecular Simulation 5, 353 (1991)

- Wolfram S., Theory and Applications of Cellular Automata. World Scientific (1986)

- Wong C. F., and McCammon J. A., Dynamics and Design of Enzymes and Inhibitors. J. Am. Chem. Soc. 108, 3830-3832 (1986)

- Wüthrich K., NMR of proteins and nucleic acids. Wiley, New York (1986)

- Yewdall S. J., Structural studies of β-lactoglobulin. PhD. Thesis, Astbury Department of Biophysics, The University of Leeds (1988)

- Zuber H., Suter F., and Kayser H., The complete amino-acid sequence of the bilin-binding protein from *Pieris brassicae* and its similarity to a family of serum transport proteins like the retinol-binding proteins. Biol. Chem. Hoppe-Seyler 369, 497-505 (1987)

Appendix A

Small Molecule Data

Atom	x	y	z	U_{eq}
Br(1)	0.07512(3)	0.41559(6)	0.12033(5)	0.0654(4)
O(2)	0.19449(19)	0.0814(3)	0.5355(3)	0.073(3)
C(6B)	0.0676(4)	-0.1366(5)	0.5977(5)	0.066(5)
N(2)	0.11522(21)	0.1582(3)	0.5836(3)	0.047(3)
N(4)	0.11753(19)	0.5393(3)	0.4192(3)	0.038(3)
O(3)	0.02598(20)	0.1928(3)	0.6376(3)	0.062(3)
O(1)	0.26397(18)	0.7254(3)	0.2769(3)	0.062(3)
C(3B)	0.14588(25)	0.0824(5)	0.5640(4)	0.047(4)
C(3)	0.07634(23)	0.5936(4)	0.3441(4)	0.044(3)
N(1)	0.15834(19)	0.7001(3)	0.3427(3)	0.040(3)
C(9B)	0.0546(3)	0.0412(4)	0.6116(4)	0.043(4)
C(5B)	0.1138(3)	-0.0811(5)	0.5748(4)	0.061(4)
C(1B)	0.0600(3)	0.1382(4)	0.6144(4)	0.046(4)
C(1')	0.0997(3)	0.4449(4)	0.4090(4)	0.045(4)
C(7)	0.3116(3)	0.7349(5)	0.2210(5)	0.070(5)
C(6)	0.1988(3)	0.6514(4)	0.4218(4)	0.050(4)
C(1A)	0.17823(25)	0.7871(4)	0.3315(4)	0.039(3)
C(4B)	0.1063(3)	0.0079(4)	0.5821(4)	0.043(4)
C(8B)	0.0088(3)	-0.0122(5)	0.6335(4)	0.057(4)
C(6A)	0.1447(3)	0.8610(4)	0.3475(4)	0.052(4)
C(2)	0.09338(23)	0.6896(4)	0.3539(4)	0.046(4)
C(2')	0.1353(3)	0.3887(4)	0.4910(4)	0.048(4)
C(7B)	0.0166(4)	-0.1019(5)	0.6271(5)	0.067(5)
C(5)	0.18441(24)	0.5549(4)	0.4138(4)	0.046(4)
C(2A)	0.2342(3)	0.8006(4)	0.2970(4)	0.045(4)
C(5A)	0.1657(3)	0.9454(5)	0.3345(5)	0.061(5)
C(4')	0.1396(3)	0.2470(4)	0.5790(4)	0.052(4)
C(3')	0.1061(3)	0.2983(4)	0.4917(4)	0.046(4)
C(3A)	0.2554(3)	0.8838(5)	0.2867(5)	0.058(4)
C(4A)	0.2215(3)	0.9566(5)	0.3043(5)	0.060(5)
Ow(1)	0.0000(0)	0.2842(7)	0.2500(0)	0.097(3)

Table A–1: Atomic coordinates and isotropic thermal factors for NAN-190

Atom	x	y	z	U_{eq}
I(1)	0.3096(8)	0.16338(4)	0.50807(2)	0.1138(5)
Br(1)	0.22082(7)	0.11441(4)	0.07385(3)	0.0691(4)
N(4)	0.7983(6)	0.0846(3)	1.09255(19)	0.052(3)
O(1)	0.5658(6)	0.1510(3)	1.29679(18)	0.078(3)
N(1)	0.7857(6)	0.0727(3)	1.22341(19)	0.051(3)
C(6)	0.8165(8)	0.1565(3)	1.19344(24)	0.058(4)
N(2')	0.6577(7)	0.1667(4)	0.81739(23)	0.081(4)
C(4B)	0.4122(9)	0.1841(5)	0.5985(3)	0.076(5)
C(5B)	0.4468(10)	0.1161(5)	0.6361(3)	0.087(5)
C(2A)	0.7132(9)	0.1128(4)	1.3261(3)	0.064(4)
C(5)	0.7299(8)	0.1566(3)	1.12949(24)	0.054(3)
C(1A)	0.8317(7)	0.0724(4)	1.28773(25)	0.057(4)
C(2)	0.8694(8)	0.0055(4)	1.1886(3)	0.064(4)
C(1B)	0.5408(8)	0.2122(4)	0.7170(3)	0.065(4)
C(1')	0.7058(7)	0.0797(4)	1.02930(24)	0.063(4)
C(3)	0.7784(8)	0.0012(3)	1.1252(3)	0.062(4)
C(6A)	0.9828(9)	0.0321(5)	1.3142(3)	0.084(5)
C(3A)	0.7448(11)	0.1124(5)	1.3891(3)	0.081(5)
C(2B)	0.5049(9)	0.2801(4)	0.6782(3)	0.078(5)
C(4')	0.7418(11)	0.1822(7)	0.8790(3)	0.107(7)
O(2)	0.6381(8)	0.3065(3)	0.79708(22)	0.112(4)
C(5')	0.6167(9)	0.2325(5)	0.7809(3)	0.076(5)
C(3B)	0.4392(10)	0.2663(5)	0.6177(3)	0.085(5)
C(2')	0.7531(9)	0.1537(5)	0.9895(3)	0.080(5)
C(5A)	1.0115(11)	0.0317(6)	1.3787(4)	0.100(6)
C(4A)	0.8958(12)	0.0711(6)	1.4142(3)	0.097(6)
C(7)	0.4264(12)	0.1801(6)	1.3328(4)	0.107(6)
C(6B)	0.5092(10)	0.1302(5)	0.6958(3)	0.083(5)
C(3')	0.6561(9)	0.1390(6)	0.9256(3)	0.095(5)
Ow(1)	0.6328(6)	-0.0275(3)	0.79347(22)	0.090(3)

Table A–2: Atomic coordinates and isotropic thermal factors for IMD-1

Bond	Length(Å)	Bond	Length(Å)
Br(1)-Ow(1)	3.349(6)	C(1')-C(2')	1.518(8)
Br(1)-Hw(1)	3.07(6)	O(2)-C(3B)	1.211(8)
C(6B)-C(5B)	1.405(10)	C(6B)-C(7B)	1.372(10)
N(2)-C(3B)	1.389(8)	C(6)-C(5)	1.499(8)
N(2)-C(1B)	1.399(8)	C(1A)-C(6A)	1.386(8)
N(2)-C(4')	1.457(8)	C(1A)-C(2A)	1.420(8)
N(4)-C(3)	1.492(7)	C(8B)-C(7B)	1.378(10)
N(4)-C(1')	1.487(7)	C(6A)-C(5A)	1.386(9)
N(4)-C(5)	1.501(7)	C(2')-C(3')	1.515(8)
Hn(4)-C(3)	2.116(7)	O(3)-C(1B)	1.204(7)
Hn(4)-C(1')	2.069(7)	O(1)-C(7)	1.434(8)
Hn(4)-C(5)	2.096(7)	O(1)-C(2A)	1.374(7)
C(3B)-C(4B)	1.479(8)	N(1)-C(6)	1.474(7)
C(2A)-C(3A)	1.364(9)	N(1)-C(1A)	1.411(7)
C(3)-C(2)	1.506(8)	N(1)-C(2)	1.472(7)
C(9B)-C(1B)	1.478(8)	C(5A)-C(4A)	1.384(10)
C(9B)-C(4B)	1.379(8)	C(4')-C(3')	1.513(8)
C(9B)-C(8B)	1.373(9)	C(3A)-C(4A)	1.382(9)
C(5B)-C(4B)	1.368(9)	Ow(1)-Hw(1)	0.76(6)

Table A–3: Bond lengths (Å) for non-hydrogen atoms of NAN-190

Bond	Length(Å)	Bond	Length(Å)
I(1)-C(4B)	2.105(7)	C(2)-C(3)	1.510(8)
N(4)-C(5)	1.493(7)	C(1B)-C(2B)	1.375(9)
N(4)-C(1')	1.511(7)	C(1B)-C(5')	1.510(9)
N(4)-C(3)	1.496(7)	C(1B)-C(6B)	1.377(9)
C(1')-C(2')	1.502(9)	O(1)-C(2A)	1.372(7)
O(1)-C(7)	1.416(10)	N(1)-C(6)	1.488(7)
N(1)-C(1A)	1.430(7)	C(6A)-C(5A)	1.416(11)
N(1)-C(2)	1.458(7)	C(3A)-C(4A)	1.380(11)
C(6)-C(5)	1.506(8)	C(2B)-C(3B)	1.400(10)
N(2')-C(4')	1.473(10)	N(2')-C(5')	1.325(9)
C(4')-C(3')	1.408(11)	O(2)-C(5')	1.217(9)
C(2')-C(3')	1.553(10)	C(5A)-C(4A)	1.343(12)
C(4B)-C(5B)	1.358(10)	C(4B)-C(3B)	1.362(10)
C(5B)-C(6B)	1.379(10)	C(2A)-C(1A)	1.404(8)
C(2A)-C(3A)	1.387(9)	C(1A)-C(6A)	1.385(9)

Table A–4: Bond lengths (Å) for non-hydrogen atoms of IMD-1

	Angle(°)		Angle(°)
C(5B)-C(6B)-C(7B)	120.5(7)	C(6A)-C(1A)-C(2A)	117.4(5)
C(3B)-N(2)-C(1B)	111.3(5)	C(3B)-C(4B)-C(9B)	108.4(5)
C(3B)-N(2)-C(4')	124.3(5)	C(3B)-C(4B)-C(5B)	131.4(6)
C(1B)-N(2)-C(4')	124.3(5)	C(9B)-C(4B)-C(5B)	120.3(6)
C(3)-N(4)-C(1')	110.8(4)	C(9B)-C(8B)-C(7B)	117.8(6)
C(3)-N(4)-C(5)	109.7(4)	C(7)-O(1)-C(2A)	117.2(5)
C(1')-N(4)-C(5)	112.8(4)	C(1A)-C(6A)-C(5A)	121.9(6)
N(1)-C(1A)-C(2A)	118.8(5)	C(1A)-N(1)-C(2)	116.1(4)
N(1)-C(6)-C(5)	110.5(5)	C(1B)-C(9B)-C(4B)	108.0(5)
N(1)-C(1A)-C(6A)	123.8(5)	C(1B)-C(9B)-C(8B)	129.9(6)
O(2)-C(3B)-N(2)	124.6(6)	C(6)-N(1)-C(1A)	113.7(4)
O(2)-C(3B)-C(4B)	129.2(6)	C(6)-N(1)-C(2)	108.3(4)
N(2)-C(3B)-C(4B)	106.2(5)	C(6B)-C(7B)-C(8B)	121.2(7)
C(3)-C(2)-N(1)	108.8(4)	C(1')-C(2')-C(3')	111.2(5)
N(4)-C(3)-C(2)	111.5(4)	C(4B)-C(9B)-C(8B)	122.1(6)
C(6B)-C(5B)-C(4B)	118.2(6)	N(4)-C(5)-C(6)	110.0(5)
N(2)-C(1B)-O(3)	123.7(5)	N(4)-C(1')-C(2')	112.7(4)
N(2)-C(1B)-C(9B)	106.2(5)	C(1A)-C(2A)-C(3A)	120.4(5)
O(3)-C(1B)-C(9B)	130.2(6)	C(6A)-C(5A)-C(4A)	119.4(6)
O(1)-C(2A)-C(1A)	115.3(5)	O(1)-C(2A)-C(3A)	124.3(5)
N(2)-C(4')-C(3')	113.0(5)	C(2A)-C(3A)-C(4A)	121.1(6)
C(2')-C(3')-C(4')	109.6(5)	C(5A)-C(4A)-C(3A)	119.7(6)

Table A–5: Angles (°) for non-hydrogen atoms in NAN-190

	Angle(°)		Angle(°)
C(5B)-C(4B)-C(3B)	122.0(7)	C(4B)-C(5B)-C(6B)	119.4(7)
C(5)-N(4)-C(1')	112.6(4)	C(5)-N(4)-C(3)	110.4(4)
O(1)-C(2A)-C(1A)	115.3(5)	C(1')-N(4)-C(3)	110.0(4)
O(1)-C(2A)-C(3A)	123.4(6)	C(1A)-C(2A)-C(3A)	121.3(6)
N(4)-C(5)-C(6)	111.3(4)	C(2A)-O(1)-C(7)	118.0(5)
C(6)-N(1)-C(1A)	113.8(4)	C(6)-N(1)-C(2)	108.6(4)
C(1A)-N(1)-C(2)	115.5(4)	N(1)-C(1A)-C(2A)	117.9(5)
N(1)-C(6)-C(5)	110.0(4)	N(1)-C(1A)-C(6A)	123.6(5)
C(2A)-C(1A)-C(6A)	118.5(5)	N(1)-C(2)-C(3)	109.4(5)
C(4')-N(2')-C(5')	119.7(6)	C(2B)-C(1B)-C(5')	117.4(6)
C(2B)-C(1B)-C(6B)	119.0(6)	C(5')-C(1B)-C(6B)	123.6(6)
N(4)-C(1')-C(2')	112.6(5)	N(4)-C(3)-C(2)	110.4(5)
C(1A)-C(6A)-C(5A)	119.3(6)	I(1)-C(4B)-C(5B)	119.7(5)
C(2A)-C(3A)-C(4A)	118.8(7)	I(1)-C(4B)-C(3B)	118.3(5)
C(3A)-C(4A)-C(5A)	121.2(8)	C(1B)-C(2B)-C(3B)	120.7(6)
N(2')-C(4')-C(3')	113.5(7)	C(5B)-C(6B)-C(1B)	120.6(6)
N(2')-C(5')-C(1B)	117.2(6)	N(2')-C(5')-O(2)	122.5(7)
C(4')-C(3')-C(2')	112.1(7)	C(1B)-C(5')-O(2)	120.3(6)
C(4B)-C(3B)-C(2B)	118.3(7)	C(1')-C(2')-C(3')	107.3(5)
C(6A)-C(5A)-C(4A)	120.8(8)		

Table A–6: Angles (°) for non-hydrogen atoms in IMD-1

	Angle(°)		Angle(°)
C(7B)-C(6B)-C(5B)-C(4B)	-0.3(10)	C(1')-N(4)-C(5)-C(6)	178.5(4)
C(5B)-C(6B)-C(7B)-C(8B)	1.0(11)	C(1B)-N(2)-C(3B)-O(2)	179.4(6)
C(1B)-N(2)-C(3B)-C(4B)	1.2(6)	C(4')-N(2)-C(3B)-O(2)	-4.5(9)
C(4')-N(2)-C(3B)-C(4B)	177.2(5)	C(3B)-N(2)-C(1B)-O(3)	178.4(6)
C(3B)-N(2)-C(1B)-C(9B)	-1.6(6)	C(4')-N(2)-C(1B)-O(3)	2.4(9)
C(4')-N(2)-C(1B)-C(9B)	-177.7(5)	C(3B)-N(2)-C(4')-C(3')	107.1(6)
C(1B)-N(2)-C(4')-C(3')	-77.4(7)	C(1')-N(4)-C(3)-C(2)	179.4(4)
C(5)-N(4)-C(3)-C(2)	-55.4(6)	C(7)-O(1)-C(2A)-C(1A)	165.6(5)
C(7)-O(1)-C(2A)-C(3A)	-15.1(8)	O(2)-C(3B)-C(4B)-C(9B)	-178.3(6)
C(3)-N(4)-C(1')-C(2')	-173.2(4)	O(2)-C(3B)-C(4B)-C(5B)	2.0(11)
N(2)-C(3B)-C(4B)-C(9B)	-0.2(6)	N(2)-C(3B)-C(4B)-C(5B)	-179.8(6)
N(4)-C(3)-C(2)-N(1)	59.7(6)	C(5)-N(4)-C(1')-C(2')	63.4(6)
C(3)-N(4)-C(5)-C(6)	54.4(6)	C(1A)-N(1)-C(6)-C(5)	-166.0(5)
C(2)-N(1)-C(6)-C(5)	63.3(6)	C(2A)-C(1A)-C(6A)-C(5A)	-2.6(9)
C(6)-N(1)-C(1A)-C(6A)	-114.8(6)	N(1)-C(1A)-C(2A)-O(1)	0.4(8)
C(6)-N(1)-C(1A)-C(2A)	68.6(6)	N(1)-C(1A)-C(2A)-C(3A)	-178.9(5)
C(2)-N(1)-C(1A)-C(6A)	11.9(8)	C(6A)-C(1A)-C(2A)-O(1)	-176.5(5)
C(2)-N(1)-C(1A)-C(2A)	-164.8(5)	C(6A)-C(1A)-C(2A)-C(3A)	4.2(9)
C(6)-N(1)-C(2)-C(3)	-62.3(5)	C(9B)-C(8B)-C(7B)-C(6B)	-1.5(11)
C(1A)-N(1)-C(2)-C(3)	168.4(4)	C(1A)-C(6A)-C(5A)-C(4A)	0.2(10)
C(4B)-C(9B)-C(1B)-N(2)	1.5(6)	C(4B)-C(9B)-C(1B)-O(3)	-178.6(6)
C(8B)-C(9B)-C(1B)-N(2)	-179.1(6)	C(1')-C(2')-C(3')-C(4')	-175.5(5)
C(8B)-C(9B)-C(1B)-O(3)	0.9(11)	C(1B)-C(9B)-C(4B)-C(3B)	-0.8(6)
C(1B)-C(9B)-C(4B)-C(5B)	178.9(6)	C(8B)-C(9B)-C(4B)-C(3B)	179.7(6)
C(8B)-C(9B)-C(4B)-C(5B)	-0.6(9)	C(1B)-C(9B)-C(8B)-C(7B)	-178.1(6)
C(4B)-C(9B)-C(8B)-C(7B)	1.3(10)	O(1)-C(2A)-C(3A)-H(3A)	-3.1(11)
C(6B)-C(5B)-C(4B)-C(3B)	179.7(6)	O(1)-C(2A)-C(3A)-C(4A)	177.2(6)
C(6B)-C(5B)-C(4B)-C(9B)	0.1(9)	C(1A)-C(2A)-C(3A)-C(4A)	-3.6(10)
C(6A)-C(5A)-C(4A)-C(3A)	0.6(10)	N(4)-C(1')-C(2')-C(3')	166.9(4)
N(2)-C(4')-C(3')-C(2')	-178.6(5)	N(1)-C(6)-C(5)-N(4)	-59.4(6)
C(2A)-C(3A)-C(4A)-C(5A)	1.1(10)	N(1)-C(1A)-C(6A)-C(5A)	-179.3(6)

Table A–7: Torsion angles (°) of non-hydrogen atoms for NAN-190

	Angle(°)		Angle(°)
C(1')-N(4)-C(5)-C(6)	176.5(4)	C(3)-N(4)-C(5)-C(6)	53.2(6)
C(7)-O(1)-C(2A)-C(1A)	169.6(6)	C(7)-O(1)-C(2A)-C(3A)	-9.7(9)
C(5)-N(4)-C(1')-C(2')	68.6(6)	C(1A)-N(1)-C(6)-C(5)	-168.3(4)
C(3)-N(4)-C(1')-C(2')	-167.9(5)	C(2)-N(1)-C(6)-C(5)	61.6(5)
C(6)-N(1)-C(1A)-C(2A)	75.6(6)	C(6)-N(1)-C(1A)-C(6A)	-106.7(6)
C(2)-N(1)-C(1A)-C(2A)	-157.8(5)	C(2)-N(1)-C(1A)-C(6A)	20.0(8)
C(5)-N(4)-C(3)-C(2)	-54.8(6)	C(6)-N(1)-C(2)-C(3)	-63.4(5)
C(1')-N(4)-C(3)-C(2)	-179.7(4)	C(1A)-N(1)-C(2)-C(3)	167.5(4)
N(1)-C(6)-C(5)-N(4)	-56.7(6)	C(5')-N(2')-C(4')-C(3')	130.1(7)
C(4')-N(2')-C(5')-C(1B)	176.5(6)	C(4')-N(2')-C(5')-O(2)	-3.4(11)
C(5')-C(1B)-C(2B)-C(3B)	178.2(6)	C(6B)-C(1B)-C(2B)-C(3B)	-0.8(10)
C(2B)-C(1B)-C(5')-N(2')	-177.4(6)	C(2B)-C(1B)-C(5')-O(2)	2.5(10)
C(6B)-C(1B)-C(5')-N(2')	1.5(10)	C(6B)-C(1B)-C(5')-O(2)	-178.6(7)
C(2B)-C(1B)-C(6B)-C(5B)	1.7(10)	C(5')-C(1B)-C(6B)-C(5B)	-177.2(6)
I(1)-C(4B)-C(5B)-C(6B)	-178.2(5)	N(4)-C(1')-C(2')-C(3')	177.7(5)
C(3B)-C(4B)-C(5B)-C(6B)	0.8(11)	I(1)-C(4B)-C(3B)-C(2B)	179.0(5)
C(5B)-C(4B)-C(3B)-C(2B)	0.1(11)	C(4B)-C(5B)-C(6B)-C(1B)	-1.7(11)
C(1A)-C(6A)-C(5A)-C(4A)	-1.1(12)	O(1)-C(2A)-C(1A)-N(1)	-1.8(8)
O(1)-C(2A)-C(1A)-C(6A)	-179.6(5)	C(2A)-C(3A)-C(4A)-C(5A)	0.1(12)
C(3A)-C(2A)-C(1A)-N(1)	177.5(6)	C(3A)-C(2A)-C(1A)-C(6A)	-0.4(9)
O(1)-C(2A)-C(3A)-C(4A)	179.0(6)	C(1B)-C(2B)-C(3B)-C(4B)	-0.1(10)
C(1A)-C(2A)-C(3A)-C(4A)	-0.2(10)	N(1)-C(1A)-C(6A)-C(5A)	-176.7(6)
N(2')-C(4')-C(3')-C(2')	177.6(6)	C(2A)-C(1A)-C(6A)-C(5A)	1.0(10)
N(1)-C(2)-C(3)-N(4)	60.7(6)	C(1')-C(2')-C(3')-C(4')	-156.3(6)
C(6A)-C(5A)-C(4A)-C(3A)	0.6(13)		

Table A–8: Torsion angles (°) of non-hydrogen atoms for IMD-1